VNIVERSITAT DE VALÈNCIA

Programa de doctorado en ESTADÍSTICA Y OPTIMIZACIÓN

SPATIO-TEMPORAL ANALYSIS OF INFECTIOUS DISEASES

Realizada por: Daniel Adyro Martínez Bello

Director: Profesor Antonio López Quílez

Noviembre, 2018

D. Antonio López Quílez, profesor titular, Departamento de Estadística e Investigación Operativa de la Universidad de Valencia

CERTIFICA que la presente memoria de investigación

"SPATIO-TEMPORAL ANALYSIS OF INFECTIOUS DISEASES"

ha sido realizada bajo su dirección por Daniel Adyro Martínez Bello, y constituye su tesis para optar al grado de Doctor en Estadística y Optimización.

Y para que así conste, en cumplimiento con la normativa vigente, autoriza su presentación ante la Facultad de Matemáticas de la Universidad de Valencia para que pueda ser tramitada su lectura y defensa pública.

Valencia, 20 de noviembre de 2018.

Antonio López Quílez

A mi padres, hermanos e hija

Agradecimientos

Los trabajos que se presentan en esta memoria de tesis de doctorado corresponden a los productos del proceso de formación académica y crecimiento personal llevado a cabo en un período de cerca de siete años, en el cual han intervenido personas a las cuales estoy profundamente agradecido, y quienes seran presentadas a continuación.

Agradezco a mi Director de Tesis, el Profesor de la Facultad de Matemáticas de la Universidad de Valencia Antonio López Quílez, quien me ha acompañado en este proceso, ha sido mi mentor, y el guía en esta empresa de desarrollar un proyecto de tesis doctoral. Recuerdo una de las primeras reuniones, a la cual llegue, presenté mi material, y procedí a despedirme, y el me dijo: *ey!, para donde vas que apenas empezamos*. Las reuniones de discusión de hallazgos de investigación han sido las vivencias mas interesantes e inolvidables, y han sido el corazón de toda la experiencia de ser un estudiante de doctorado. Estas reuniones me enseñaron la importancia de compartir ideas y diferentes puntos de vista cuando se enfrenta un problema de investigación.

Quiero agradecer a mi hermano, el Profesor de la Facultad de Magisterio de la Universidad de Valencia Vladimir Martínez Bello, quien con su alegría, optimismo y energía vital ha sido el gestor de mi proceso de formación.

También quiero agradecer a Alexander Torres Prieto, coordinador de la Oficina de Epidemiología y Demografía de la Secretaría de Salud del Departamento de Santander, quien colaboró en este proyecto de investigación de las enfermedades del dengue y del virus del Zika en Colombia.

Mi agradecimiento al estado de la República de Colombia, que financió mi estudio doctoral en la Universidad de Valencia a traves de la convocatoria 646 de 2014 del Departamento Administrativo de Ciencia y Tecnología (COLCIENCIAS).

También debo agradecer a la Profesora María Dolores Ugarte y al Profesor Aritz Adin, en la Universidad Pública de Navarra, quienes tuvierón la buena disposición para colaborar en el desarrollo de uno de los trabajos de investigación presentados en esta memoria.

I would also like to thank Professor Ziv Shkedy in Hasselt University, Belgium, who extended the invitation to make the research internship in the Center for Statistics at Hasselt University, and the research collaboration leading to one of the publications shown in the present Thesis.

Gratitud inmensa a mi cuñada preferida, Angela Martínez, quien me acompaño durante el tiempo de desarrollo de este trabajo con su buena energía y con las mejores lentejas que se puedan degustar al oeste del *mare nostrum*.

Agradezco a mis padres, el Doctor Jose Martínez Cortes y la Licenciada Zorayda Bello de Martínez, quienes me han apoyado incansablemente a lo largo de toda mi existencia en todos los proyectos de vida que he afrontado.
Agradezco con mucho amor, a la señorita Sara Gabriela, quien ha sido mi inspiración para llevar a cabo este trabajo.
Finalmente, quiero agradecer a la vida que me ha dado tanto.

Organización de la tesis/ Thesis organization

De acuerdo con la normativa vigente de la Universitat de València (Reglament sobre depòsit, avaluació i defensa de la tesi doctoral, aprovat per el Consell de Gobern de 28 de juny de 2016, Article 8) esta Tesis Doctoral se presenta como compendio de publicaciones. El Capítulo 1 corresponde a un resumen de la Tesis que incluye introducción, metodología, principales resultados y conclusiones. Los Capítulos 2 a 9 corresponden a ocho artículos elaborados. Seis artículos están ya publicados, uno está enviado a revisión y el otro en fase final de preparación para ser sometido a publicación. En todos ellos como primer autor, salvo dos como segundo autor. Todos los artículos han sido enviados o van a serlo pronto a revistas indexadas en Journal Citation Reports (JCR), con altos factores de impacto. A continuación se presenta el listado de las publicaciones presentadas en esta Tesis:

Following the current regulations of Universitat de València (Reglament sobre depòsit, avaluació i defensa de la tesi doctoral, aprovat per el Consell de Gobern de 28 de juny de 2016, Article 8) this Doctoral Thesis is presented as a compendium of publications. Chapter 1 corresponds to the Thesis summary including introduction, methodology, main results and conclusions. Chapters 2 to 9 correspond to eight prepared articles. Six articles have already been published, one has been sent to peer-review and the other is in the final phase of preparation to be submitted to publication. In all articles as first author, except two articles as second author. All articles have been or will soon be sent to indexed journals in Journal Citation Reports (JCR), with high impact factors. The list of publications is presented next:

- **Martínez-Bello DA**, López-Quílez A, Torres Prieto A. Bayesian dynamic modeling of time series of dengue disease case counts. *PLoS Neglected Tropical Diseases*. 2017, 11(7): e0005696.
- **Martínez-Bello DA**, López-Quílez A, Torres Prieto A. Relative risk estimation of dengue disease at small spatial scale. *International Journal of Health Geographics*. 2017, 16:31.
- **Martínez-Bello DA**, López-Quílez A, Torres Prieto A. Spatiotemporal modeling

of relative risk of dengue disease in Colombia. *Stochastic Environmental Research and Risk Assessment*. 2017, 32(6): 1587.

- Adin A, **Martínez-Bello DA**, López-Quílez A, Ugarte MD. Two-level resolution of relative risk of dengue disease in a hyperendemic city of Colombia. *PLoS ONE*. 2018, 13(9): e0203382.

- **Martínez-Bello DA**, López-Quílez A, Torres Prieto A. Spatio-temporal modeling of Zika and dengue infections within Colombia. *International Journal of Environmental Research and Public Health*. 2018, 15: 137.

- **Martínez-Bello DA**, López-Quílez A, Torres Prieto A. Joint estimation of relative risk of dengue and Zika Infections. *Manuscript submitted after revision.*

- Sebrango-Rodríguez CR, **Martínez-Bello DA**, Sánchez-Valdés L, Thilakarathne PJ, Del Fava E, Vand Der Stuyft P, López-Quílez A, Shkedy Z. Real time parameter estimation of Zika outbreaks using model averaging. *Epidemiology and Infection*. 2017, 145: 2313–2323.

- **Martínez-Bello DA**, López-Quílez A, Shkedy Z. Bayesian Nonlinear Models for Estimation and Prediction for Zika outbreak data. *Manuscript in preparation.*

Resumen

Los sistemas de vigilancia de salud pública colectan y analizan datos que soportan los programas de control y prevención de enfermedades en todo el mundo. En Colombia, el sistema de vigilancia en salud pública (SIVIGILA) esta encargado del flujo de datos e información de la vigilancia de las enfermedades de notificación obligatoria que afectan la salud de los Colombianos. Las enfermedades transmitidas por mosquitos tales como el dengue, la malaria, la fiebre amarilla, la enfermedad del virus del Chikungunya, la enfermedad del virus del Zika (EVZ) entre otras afectan seriamente la salud de las poblaciones a través de todo el país. Dentro de estas enfermedades se destacan la enfermedad del dengue y la EVZ. El dengue es responsable de una gran cantidad de personas enfermas con algunos casos de mortalidad, desde la decada de los ochenta en el siglo veinte, mientras que la EVZ se ha reportado en el país desde el segundo semestre del año 2015 asociada a severos síndromes neurológicos en neonatos y adultos.

En esta tesis por compendio de publicaciones se exploran métodos estadísticos jerárquicos Bayesianos para la evaluación del riesgo espacial y temporal del dengue y la EVZ en varios niveles de agregación temporal y espacial de los datos post-procesados del sistema de vigilancia en Colombia, especialmente motivados por explorar los problemas y desafíos de la implementación de estos modelos.

La estructura de la tesis consiste de un capítulo introductorio, y ocho capítulos que corresponden a un número igual de artículos de investigación. El capítulo uno es un resumen general de la disertación que presenta los objetivos, la metodología, resultados y conclusiones del trabajo de investigación. El segundo capítulo analiza datos temporalmente agregados de casos de dengue y covariables meteorológicas asociadas a la enfermedad utilizando modelos con parámetros que varian en el tiempo. El capítulo tres estudia modelos espaciales de riesgo de dengue con parámetros que varian en el espacio y covariables derivadas de datos de satélite a nivel de ciudad. El capítulo cuatro explora modelos espacio-temporales de riesgo de dengue incluyendo covariables derivadas de datos de satélite con parámetros que varian en el tiempo a nivel de ciudad. El capítulo cinco desarrolla modelos espacio-temporales de varios niveles geográficos de agregación para la estimación de riesgo de dengue a nivel de ciudad. El sexto capítulo desarrolla la estimación del riesgo en paralelo de dengue y la EVZ a nivel de ciudad y de departamento. El capítulo siete desarrolla la estimación de riesgo conjunto de dengue y EVZ utilizando modelos multivariados jerárquicos Bayesianos a nivel de ciudad y de departamento. La estimación de los parámetros de los modelos en los capítulos dos, tres, cuatro y siete se desarrolla usando métodos de Monte Carlo de Cadenas de Markov, mientras que lo capítulos cinco y seis utilizan “integrated nested Laplace approximation” (INLA). Los capítulos ocho y nueve presentan modelos no-lineales para los datos acumulados de los casos de EVZ en diferentes ciudades de Colombia. El capítulo ocho realiza la estimación de los parámetros por medio del método de mínimos cuadrados no-lineales, mientras que el capítulo nueve utiliza Monte Carlo Hamiltoniano para el mismo objetivo.

Abstract

Public health surveillance systems collect and analyze data supporting programs of controlling and preventing diseases all around the world. In Colombia, the public health surveillance system (SIVIGILA) is in charge of the data and information flow for the surveillance of obligatory notification diseases affecting the Colombian population health. Diseases transmitted by mosquitoes such as dengue, malaria, yellow fever, Chikungunya fever, Zika virus disease (ZVD) among other seriously affect health populations along the country. Within these diseases, dengue and ZVD are highlighted. Dengue is responsible of a great burden of sick people with some cases of mortality since the eighties in the twenty century, while ZVD has been reported in the country since the second semester of year 2015 associated to severe neurological syndrome in newborns and adults.
In this thesis by compendium of publications are explored hierarchical Bayesian statistical methods for the assessment of temporal and spatial dengue and ZVD risk at some temporal and spatial aggregation level using post-processed data obtained from the surveillance system in Colombia, specially motivated by exploring model implementation problems and challenges.
The dissertation structure consist in one introductory chapter, and eight chapters corresponding to an equal number of research papers. Chapter one is an overall summary of the dissertation presenting the objectives, methodology, results and conclusions of the research work. The second chapter analyzes temporally aggregated data of dengue and meteorological covariates associated with the disease using dynamic models with time-varying parameters. The chapter three studies spatial models of dengue risk with space-varying parameters and covariates derived from satellite data at city-level. The chapter four explores spatio-temporal models of dengue risk including covariates derived from satellite data with time-varying parameters. The chapter five develops spatio-temporal models of dengue risk at two geographic levels of aggregation at city-level. The chapter six develops the parallel estimation of dengue and ZVD risk at departmental and city level. The chapter seven develops the joint estimation of dengue and ZVD risk using hierarchical Bayesian multivariate models at departmental and city level. Parameter estimation in chapters two, three, four, and seven are developed using Monte Carlo Markov Chain methods, while chapters five and six used ”integrated nested Laplace approximation” (INLA). The chapters eight and nine present nonlinear models for the cumulative data of the ZVD cases in several Colombian cities. The chapter eight makes parameter estimation by means of the nonlinear least squares, while chapter nine presents Hamiltonian Monte Carlo for the same objective.

Contents

List of Figures

List of Tables

Chapter 1

Thesis digest

1.1 Introduction

Human vector-borne diseases are disseminated in tropical and subtropical regions of the world. They cause great amount of expenses represented in medical services, incapacities and death. Only in the Americas, dengue and yellow fever are viral diseases transmitted by mosquitoes from the genus Aedes affecting the health of 2,276,803 and 111 people respectively during 2016 [1]. Public health surveillance systems in South America collect event data, generating information for decision making in public health. In Colombia, the public health surveillance system (SIVIGILA for the Spanish acronym) is maintained by the Colombian National Health Institute (INS), covers near 11,000 health services providers, plus 1117 municipalities, 32 departments and 5 districts. SIVIGILA provides protocols for the events of public health interest (EPHI) to be notified, user manuals, data dictionaries, notification forms of near to 76 events, and giving feedback to more than 114 public and private health insurance companies. The objectives of SIVIGILA are to systematically collect, analyze, interpret, update, publish, and evaluate the data related with health events for the orientation of prevention and control activities in public health [2]. Colombia's 2016 population is nearly 48 millions people, living in 33 administrative units known as departments. The department of Santander is located to the north eastern of Colombia, and its capital is the city of Bucaramanga with 521,000 habitants. Particularly, the population of Bucaramanga has been highly affected by dengue disease in the last ten years. In addition, since 2015, population in South America has been highly affected by the Zika virus disease (ZVD) (other viral disease transmitted by *Aedes* mosquitoes).
The main motivation of the present Thesis was the observation that data for arboviral diseases such as dengue and ZVD collected for the SIVIGILA have been used to produce information related with disease events by the production of municipal, departmental and national public health reports, with detailed epidemiological information, but a superficial use of spatial statistical and epidemiological models.
Spatial statistical analysis is important to the public health due to the information contained in spatial and temporal scales, and the knowledge acquired by including multi-

factorial processes in the public health decision making process. Waller and Gotway [10] state that spatial statistical methods contribute to evaluate differences in rates observed from different geographic regions, separate pattern from background noise, identify clusters of events, and assess the importance of possible local or global influences.
In this Thesis by compendium of publications, we have focused but not limited to Bayesian hierarchical models for the analysis of temporal, spatial and spatio-temporal count data representing cases of dengue and ZVD at several geographic and temporal aggregation scales in Colombia. Chapter 2 introduces the hierarchical Bayesian modelling of time series of dengue disease at city level. A spatial risk analysis of dengue diseases together with the inclusion of covariates obtained from satellite data at city level is covered in Chapter 3. We assess the spatio-temporal risk analysis of dengue at city level, using Markov Chain Monte Carlo (MCMC) methods in Chapter 4. In Chapter 5, we review the spatio-temporal risk analysis of dengue, including models for multiple levels of spatial aggregation, and using the integrated nested Laplace approximation (INLA) for parameter estimation. A parallel spatio-temporal risk analysis of dengue and ZVD is the focus of Chapter 6. In Chapter 7, we investigate the spatial joint risk analysis of dengue and ZVD at city and departmental level. Nonlinear models for the analyses of time series of ZVD cases are presented in Chapter 8, while Chapter 9 extends the nonlinear models for ZVD to the Bayesian hierarchical framework.
The aim of this chapter is to provide an overall summary of the dissertation. Section 1.2 provides general information about the arboviral diseases under research, a review of statistical models for spatio-temporal data of dengue and ZVD, the data sources, the general objective, and the Thesis outline. Section 1.3 describes the statistical models and estimation tools applied to the data. Section 1.4 presents an overview of the results, while Section 1.5 exposes the concluding remarks and future research.

1.2 Motivation

1.2.1 Dengue Disease

Dengue is an arboviral disease caused by a *Flavivirus* genus in the family *Flaviviridae*. Dengue virus is genetically related to the Japanese Encephalitis Virus, Yellow Fever Virus, West Nile Virus, and Murray Valley encephalitis. Dengue virus is a zoonotic arbovirus, having a sylvatic cycle of non-human primates. There are four dengue serotypes, (DENV 1-4). Dengue virus is transmitted in a urban cycle primarily by the mosquito *Aedes aegypti*, and by the *Aedes albopictus* in smallest extent [3]. Clinical manifestations of dengue range from the asymptomatic infection through a influenza-like febrile illness to severe systemic signs including capillary leak, shock, hemorrhage and death. The proportion of cases progressing to the fatal outcomes is low, while the ratio of asymptomatic to symptomatic dengue disease is approximately 4 to 1 [3]. In the Americas, dengue disease has an endemo-epidemic pattern with outbreaks every three to five years , with several outbreaks observed during the period 2011-2017 [4]. The average annual dengue

incidence in the Americas was 113, 120.7, 245, 194, 245 and 219 cases/100,000 people respectively for years 2011-2016 [4].
In Colombia, Villar et al [5] reviewed dengue epidemiology up to 2011. The dengue incidence rate was 169.7, 624.0, 122.8, 221.9, 469.9, 404.2 and 293.6 cases/100,000 people for years 2009-2015 (Source: the authors based on annual dengue reports of the Colombian National Health Institute, 2009-2015). In 2016, it was reported a total number of 103,822 dengue cases, with an incidence rate of 213.39 cases per 100,000 inhabitants. [6].

1.2.2 Zika Virus Disease

Zika is a *Flavivirus* in the *Flaviviridae* family, transmitted by mosquitoes, mainly of the *Aedes* genus, although other non-vector modes of transmission has been recognized, such as the congenital, perinatal, sexual, blood transfusions, animal bite and laboratory exposure. It causes the ZVD, characterized by rash, fever, arthralgia, myalgia, fatigue, headache and conjunctivitis, although likely 80% of cases remain asymptomatic, and rarely produces death. ZVD has been linked to central nervous system anomalies in adults, and congenital neurological malformations in children born to pregnant women infected in early stages of pregnancy [7].
ZVD was first reported in continental South America in Brazil in May 2015 [7], while 49 other countries in the Americas have reported ZVD imported and autochthonous transmission [8]. Pan American Health Organization reports 1,00,140 cases and 2530 confirmed congenital syndrome associated with ZVD for the year 2016, with an incidence rate for the Americas of 71.38 cases/100,000 people [8].
In Colombia, ZVD was first declared in early October 2015, after the first cluster of laboratory-confirmed cases was identified in nine patients from northern Colombia [9]. From the 32 epidemiological week 2015 (October) to the 52 epidemiological week 2016 (December) there were notified 106,659 ZVD cases (9.2% confirmed, and 90.8% suspected by clinical signs), with an incidence rate for the period of 219.22 cases/100,000 people [6].

1.2.3 Spatio-temporal models for dengue disease data

Racloz et al. [11] reviewed epidemiological tools for dengue surveillance, identifying mapping tools, mathematical models and the combination of both, where these techniques were aimed at dengue reporting surveillance usually based on risk factors, and some reports introducing disease forecasting. Racloz et al. [11] found research reports combining serological, climate, household and community, geographical data and socio-economic information with temporal and spatial dengue data, as input into Autoregressive Moving Average (ARIMA) models, Poisson regression models, time series models, multiple regression, logistic regression, classification and regression trees (CART) and spatio-temporal regression models.
Louis et al. [12] collected and described the available dengue risk maps, the modeling

predictors used for producing the risk maps, and the methods for dengue risk mapping, discussing their impact for public health. All the reviewed studies were retrospective, using dengue data from surveillance systems. In addition to the outcome corresponding to reported dengue cases, predictors were used from a mix of categories such as population, demography, socioeconomic status, climate, environment and entomological data. Louis et al. [12] reported the following methods for dengue risk mapping: spatial analysis of case clusters/hotspots, spatial autocorrelation measures, logistic regression and multinomial models, generalized linear models, generalized additive models, kernel estimation, environmental niche modeling (species distribution modeling), maximum entropy, geographically weighted regression, kriging and co-kriging, and Knox test (space and time distance), temporal indices, and water-associated disease index (WADI).
Naish et al. [13] reviewed the association of dengue disease and climate with a focus on quantitative models of the impact of climate change. The interest was to explore the methodological issues of the association of climate factors as predictors of dengue disease. The studies included data obtained from local and national meteorological stations, and data on socio-economic heterogeneity, climatic diversity, and un-observed confounders such as social behavior. The methodological approaches were based on Seasonal Autoregressive Integrated Moving Average (SARIMA) time series, wavelet time series models, different types of regression analysis (Normal, Poisson and logistic regression), Bayesian conditional autoregressive (CAR) spatial models, generalized additive (GAM) mixed models, non-linear models, multivariate models, and global circulation models (GCM).
Runge-Ranzinger et al. [14] reviewed the use of tools for dengue outbreak prediction, and trend monitoring in passive and active disease surveillance systems. They found three main types of surveillance tools. First, surveillance systems for outbreak detection and prediction (using electronic event-/search query-based surveillance for early detection of increased dengue activity, using the appearance of a new dengue serotype/genotype as an alert signal for dengue outbreaks, using syndromic surveillance to create alert signals for dengue outbreaks, or finally using other sentinel site-based approaches to increase capacity for outbreak detection). Second, surveillance systems for describing endemic/epidemic trends (mostly population-based and passive, some included additional sentinel sites or virus surveillance but they mainly used to monitor viral trends and were not applied to early warning). Third, dengue surveillance systems in low/non-endemic countries.
Imai et al. [15] described time series models applied to infectious diseases such as dengue, malaria, cholera and influenza. The review focused on studies associating infectious diseases and environmental factors based on generalized linear models and generalized additive models, without considering the time-series methods developed from econometrics, such as the ARIMA models. The regression models included GLM (Poisson, quasi-Poisson and negative binomial), GAM (Poisson and negative binomial) and generalized mixed models, controlling seasonality and long term trend, autocorrelation (including parameters to control autocorrelation) and lag effects of exposure (lag effects of weather variables).

1.2.4 Spatio-temporal models for Zika virus disease data

Mathematical models for dynamics of ZVD have been evaluated by Wiratsudakul et al [16], however there are not extensive reviews of methodological approaches to the risk assessment of ZVD. We identify some research reports using statistical models of spatial and temporal data of ZVD.

Krystosik et al. [17] applied point process models to associate hot-spots of dengue, ZVD and Chikungunya infections with community infrastructure in a hyper-endemic Colombian city.

Riou et al. [18] analyzed the ZVD and Chikungunya epidemics occurring between 2013 and 2016 in six islands or small archipelagos of French Polynesia and three islands of the French West Indies. They fit weekly incidence data with a common hierarchical transmission model. The model has two components: (1) a mechanistic reconstruction of the distribution of the serial interval of the diseases (the time interval between disease onset in a primary and secondary case), including the influence of temperature; and (2) a Time-Series infected resistance TSIR model for the generation of observed secondary cases, that included coefficients depending on the territory, the disease and the local weather conditions.

Li et al. [19] mapped the risk of transmission of ZVD transmission in Guangdong, China, applying analytic hierarchy process (AHP), a methodology using quantitative and qualitative criteria for analyzing decision problems. AHP uses utility-based methods for multi-attribute decision-making, where objective mathematics process the subjective and personal preferences in decision making. AHP has been used in risk assessment of emerging infectious diseases (EIDs) including Dengue.

Chien et al. [20] proposed a distributed lag nonlinear model to estimate the association between the number of ZVD cases counts and meteorological measurements (humidity, rainfall, temperature among others) for weekly data in Colombia in 2015-2016. The method provides risk maps of ZVD for the 32 departments of Colombia.

Leta et al. [21] map the global risk of dengue, ZVD, yellow fever, Chikungunya fever and Rift valley fever by identifying areas where the diseases are reported, either through active transmission or travel-related outbreaks, as well as areas where the diseases are not currently reported but are nonetheless suitable for the vectors (*Aedes aegypti* and *Aedes albopictus*). This study combines the habitat suitability models of *A. aegypti* and *A. albopictus* and the disease occurrence maps to generate the risk maps.

Because the dengue and ZVD vectors also transmits urban yellow fever, we report the estimation of the yellow fever distribution and the potential spread into new areas from Shearer et al. [22]. Authors used the Poisson point process boosted regression tree model incorporating environmental and biological explanatory covariates, vaccination and spatial variability of yellow fever, to predict the relative risk of apparent yellow fever virus infection at a 5 x 5 km resolution across all risk zones (47 countries across the Americas and Africa).

Aguiar et al. [23] used a statistical Maxent model to assess the potential spatial risk of ZVD and Chikungunya disease dissemination, using ZVD and Chikungunya data in 2015 and 2016, along with environmental variables and social indicators.

1.2.5 Objective

The general objective of this Thesis is to explore spatio-temporal statistical models of areal data for the risk assessment of dengue and ZVD for some levels of temporal and geographical aggregation in Colombia.

1.2.6 Thesis outline

Throughout the Thesis, we will focus on temporal, spatial and spatio-temporal statistical models of arboviral infectious diseases.
Chapter 2 covers the analysis of the association between the weekly time series of dengue cases count with meteorological variables such as rainfall, solar radiation, relative humidity and temperature in the city of Bucaramanga, Colombia, for a period of time from January 2008-December. We apply Bayesian dynamic Poisson models, using Monte Carlo Markov chain methods for inference. It is a current practice to model time series using autoregressive integrated moving average (ARIMA) models. Instead of ARIMA models, we applied models with time-varying parameters. The content of this Chapter corresponds to the published paper in [24].
In Chapter 3, we analyze the association between dengue disease and satellite data of normalized density vegetation index (NDVI) and land surface temperature (LST) using disease mapping models in the city of Bucaramanga. We employ Poisson Normal models using structured and unstructured spatial effects, with a combination of spatially fixed and varying coefficients for the covariates, using MCMC for parameter estimation. The content of this Chapter corresponds to the published paper in [25].
The spatio-temporal modeling of relative risk of dengue disease is the focus of Chapter 4. We explore Bayesian Poisson spatio-temporal interaction effects models of relative risk of dengue disease, using data from city of Bucaramanga from January 2009 to December 2015. In addition, the association between dengue cases and environmental factors obtained from satellite images is explored using the Bayesian spatio-temporal models. The published paper in [26] corresponds to the content of this Chapter.
Chapter 5 analyzes the geographical distribution of relative risk of dengue disease in the city of Bucaramanga, and its monthly evolution in time during the period 2009-2015, identifying regional effects at two levels of areal aggregation, using the integrated nested Laplace approximation (INLA) method for parameter estimation. Chapter 5 is the product of the research collaboration with Professors María Dolores Ugarte and Aritz Adin from the Department of Statistics in the Public University of Navarra, Spain. Chapter 5 content's corresponds to the paper in [27].
Chapter 6 explores the spatio-temporal relative risk models of ZVD and dengue in parallel during the period corresponding to the ZVD outbreak in Colombia, from October 2015 to December 2016, in one high incidence city and one high incidence department of Colombia, using the epidemiological week (EW) as time measure and the census sector (city level) and municipality (departmental level) as geographic units. The published paper in [28] corresponds to the content of this Chapter.

Chapter 7 estimates the joint relative risk and the spatial association between dengue disease and ZVD infection, using data for the 2015-2016 ZVD outbreak in Colombia. We analyzed two levels of spatial data aggregation: the departmental level (disease counts aggregated per municipality) in Santander, Colombia, and the municipal level (disease counts aggregated per census section) in the city of Bucaramanga (Santander). The content of Chapter 7 is the manuscript submitted to the Emerging Infectious Diseases journal.
In Chapter 8, we analyze ZVD outbreak data and estimate a model average of the final size and the turning point of the epidemic, performing a real-time prediction using several nonlinear models. A real-time prediction is a procedure in which the final size of the outbreak is estimated as early as possible. Nonlinear models were applied to four ZVD outbreaks that occurred in four cities in Colombia during the 2015/2016 outbreak. This article was the product of the joint collaboration with Professor Ziv Shkedy in the Center for Statistics, in Hasselt University, Belgium, during the three months research stay of the author's Thesis. The published paper in [29] corresponds to the content of this Chapter.
In Chapter 9, the analysis is focused on hierarchical Bayesian nonlinear modeling and aim to conduct a real-time estimation and prediction of the final size and turning point of a disease outbreak. The proposed method is applied to outbreak data of the 2015-2016 ZVD epidemic in ten Colombian cities with the highest case counts in the outbreak. Chapter 9 corresponds to the paper in preparation for the Journal of Statistical Modelling.

1.2.7 Data sources

We used geocoded cases of dengue disease obtained from the public health surveillance system (SIVIGILA) for the period 2009-2016 for the city of Bucaramanga, Colombia, using these data in Chapter 2 to Chapter 7. We also utilized geocoded cases of ZVD for the period October 2015 to December 2016 for the city of Bucaramanga, modeling these cases in Chapter 6 and Chapter 7. In addition, we employed municipal cases of dengue and ZVD of the department of Santander, for the period October 2015 to December 2016, using these data in Chapter 6 and Chapter 7. We also modeled ZVD data obtained at municipal level, for the 10 Colombian cities with the highest incidence for the period October 2015 to December 2016, using these data in Chapter 8 and Chapter 9.
Shapefiles vector maps corresponding to the city of Bucaramanga (at census section and census sector boundaries), and the department of Santander (municipality boundaries) were obtained of the national geostatistical framework of Colombia from the National Administrative Department of Statistics (DANE). Vector maps were used in Chapter 2 to Chapter 7.
Population data by five-years age group and sex at census block level for the city of Bucaramanga and population data by five-years age group and sex at municipal level for the department of Santander were obtained from the population projections of the Colombian Census 2005, and they were employed in Chapter 2 to Chapter 9.
Meteorological data (rainfall, solar radiation, temperature and relative humidity) for the period January 2008 to December 2015 were provided by the "Corporación de Defensa

de la Meseta de Bucaramanga" (CDMB). We used these meteorological data in Chapter 2.
Satellite data of NDVI (normalized density vegetation index), LST (land surface temperature) were obtained from LANDSAT (Land Remote-Sensing Satellite System) and MODIS (Moderate Resolution Imaging Spectro-radiometer) provided by the United States Geological Service (USGS), for 2008 - 2015. Satellite data were used in Chapter 3 and Chapter 4.

1.3 Methodology

This section provides a general overview of relevant methodology applied to the research articles in the Thesis. In Subsection 1.3.1, we provide an overview of the dynamic models, such as the models with time-varying coefficients. Subsection 1.3.2 introduces basic ideas of non-linear models. The Subsection 1.3.3 introduces concepts of disease mapping, and Subsection 1.3.4 shows principles of Bayesian inference, while the Monte Carlo Markov methods for Bayesian analysis are covered in Subsection 1.3.5, and the Hamiltonian Monte Carlo in Subsection 1.3.6. Finally, Subsection 1.3.7 is devoted to the integrated nested Laplace approximation, and Subsection 1.3.8 reveals the information criteria used for model selection.

1.3.1 Dynamic models

Dynamic models are applied in Chapter 2. The dynamic linear model incorporates time varying coefficients to model time series data. The Gaussian dynamic model [30] is composed by a linear observation equation and a linear transition equation. For a sequence of data y_j $j = 1,\ldots,t$, the linear observation equation is defined by

$$y_j = \mathbf{F}_j' \boldsymbol{\theta}_j + \kappa_j \tag{1.1}$$

where $\mathbf{F}_t$ is a design vector for univariate y_j, $\kappa_j \sim \text{Normal}(0,\sigma_j^2)$ is scalar, while the linear transition equation follows

$$\boldsymbol{\theta}_j = G_j \boldsymbol{\theta}_{j-1} + \boldsymbol{\xi}_j$$

where G_j is a transition matrix, $\boldsymbol{\xi}_j \sim \text{Normal}(\mathbf{0}, Q_j)$, Q_j is a variance covariance matrix, $\boldsymbol{\theta}_0 \sim \text{Normal}(\boldsymbol{a}_0, Q_0)$ is the vector of initial values with vector mean $\boldsymbol{a}_0$ and variance covariance matrix Q_0. G_j does not depend on the time index j.
The dynamic models can be extended to non-normal data. For example, for times series of count data, a Poisson dynamic model with local level trend and two covariates with

fixed and time varying coefficient is represented by

$$\begin{aligned} y_j &\sim \text{Poisson}(\eta_j) \\ \log(\eta_j) &= \mathbf{F}_j' \boldsymbol{\theta}_j \\ &= \begin{bmatrix} 1 & x_{1,j} & x_{2,j} \end{bmatrix} \boldsymbol{\theta}_j \\ \boldsymbol{\theta}_j &= \begin{bmatrix} \theta_{1,j} \\ \theta_{2,j} \\ \theta_{3,j} \end{bmatrix} = G_j \boldsymbol{\theta}_{j-1} + \boldsymbol{\xi}_j = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_{1,j-1} \\ \theta_{2,j-1} \\ \theta_{3,j-1} \end{bmatrix} + \begin{bmatrix} u_j \\ 0 \\ v_j \end{bmatrix} \end{aligned}$$

where $x_{i,j}$ $(i = 1,2)$ are covariates and $\theta_{k,j}$ $(k = 1,2,3)$ are fixed or time-varying coefficients. $\theta_{1,j}$ is the coefficient for the first order random walk component. $\theta_{2,j}$ is a fixed coefficient and $\theta_{3,j}$ is a time-varying coefficient for one covariate. The vector $\boldsymbol{\xi}_j = [u_j, 0, v_j]'$ is assumed multivariate Normal with zero mean vector and variance-covariance matrix $\mathbf{V}$ equal to a diagonal matrix with diagonal entries $(\sigma_u^2, 0, \sigma_v^2)$.

1.3.2 Nonlinear Models

Nonlinear models are implemented in Chapter 8 and Chapter 9. Non-linear models are mechanistic, i.e, based on a model for the mechanism producing the response. The model parameters in a non-linear model have a direct physical interpretation. A non-linear model needs fewer parameters than the linear model. The non-linear models offer certain parameters summarizing characteristics of the data. The advantages of a non-linear model are parsimony, interpretability and validity beyond the range of the data [31] [32]. A general non-linear model is represented by

$$\mathbf{y} \sim \text{Normal}(f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}), \boldsymbol{\Sigma}) \tag{1.2}$$

where $f(\cdot)$ is a non-linear mean function, $\mathbf{X}$ is a covariate matrix, $\boldsymbol{\beta}$ is a vector of parameters, and $\boldsymbol{\Sigma}$ is a variance-covariance matrix.

To illustrate the nonlinear models, let us consider the four parameter logistic growth model [31] [33]. For a sequence of data y_j, $j = 1, \ldots, t$, the non-linear observation equation is defined by

$$y_j = \theta_1 + \frac{\theta_2 - \theta_1}{1 + \exp\left[\frac{(\theta_3 - x_j)}{\theta_4}\right]} \tag{1.3}$$

where θ_1 is the horizontal asymptote as x goes to infinity, θ_2 is the horizontal asymptote as x goes to minus infinite, θ_3 is the x value at the inflection point, and for the four-parameter logistic model, at this value of x the y is midway between the asymptotes, so the model is symmetric, and θ_4 is a scale parameter on the x-axis [31].

1.3.3 Disease Mapping

Disease mapping is the collection of statistical, geographical, cartographic and epidemiological tools to represent the disease in space and time. Disease mapping portraits two

characteristics, first, a spatial or geographical distribution is the focus and so the relative location of events is important, bringing the combination of geographical information systems and spatial statistics, and the second characteristics is that the interest is in the spatial distribution of disease, therefore, the main objective is how to analyze disease incidence or prevalence when we have geographical information [34]. Disease mapping is mainly associated to areal data, which is defined by Banerjee et al. [35] as *" the Y(s) random vector at location $s \in R^r$, where s varies continuously over D, and D is a fixed subset (of regular or irregular shape) partitioned into a finite number of areal units with well defined boundaries"*. The standardized incidence ratio (SIR) is the main statistic summarizing areal data for disease events [35]. It relates the observed and the expected disease values in area $i, i = 1, \ldots, n$. The expected data is a function of the incidence rate and the population by age-group, cohort, sex and many other data characteristics registered per area i. The incidence rate can be generated from the data, or can be obtained from a reference population. The first case refers to internal standardization, while the second is external standardization.

$$SIR_i = \frac{O_i}{E_i} \tag{1.4}$$

SIR are represented in choropleths maps, so the risk patterns can be distinguished. A modeling approach can be followed for the SIR, where the observed values in area i can be modeled as generated from a Poisson distribution with mean parameter equal to the expected values and the relative risk r_i in area i. The relative risk is the excess (lack) of risk of disease in area i with respect to the background risk [34].

$$O_i \sim \text{Poisson}(E_i r_i) \tag{1.5}$$

$$r_i = \exp(\eta_i) \tag{1.6}$$

The r_i is equal to the exponentiated linear predictor η_i. The linear predictor includes random effects accounting for the structured and unstructured spatial effects, and fixed and space-varying coefficients for the covariates. Many options for the linear predictor are available. In the following, we present most of the linear predictors included in this Thesis. Spatial models of relative risk are employed in Chapter 3. The starting linear predictor follows

$$\eta_i = \alpha + u_i + v_i + \sum_{k=1}^{k=K} x_{ik}\beta_k \tag{1.7}$$

$$\mathbf{u} \sim \text{Normal}(\mathbf{0}, \sigma_u^2(\mathbf{D} - \mathbf{W})^{-1}) \tag{1.8}$$

$$\mathbf{v} \sim \text{Normal}(\mathbf{0}, \sigma_v^2\mathbf{I^{-1}}) \tag{1.9}$$

$$\boldsymbol{\beta} \sim \text{Normal}(\mathbf{0}, \sigma_\beta^2\mathbf{I^{-1}}) \tag{1.10}$$

where α is the overall risk, u_i is the structured spatial effect, v_i is the unstructured spatial effect, and the β_k is a fixed effect for covariate $k = 1, \ldots, K$. The model implies the assumption of probability distributions for the coefficients of the spatial effects and covariates. First, the vector of structured spatial effects $\mathbf{u}$ is multivariate normal with

zero mean vector, and variance-covariance matrix $\sigma_u^2(\mathbf{D}-\mathbf{W})^{-1}$, where σ_u^2 is a variance parameter to be estimated from the data, and $(\mathbf{D}-\mathbf{W})^{-1}$ is a structure matrix. The structure matrix is composed by the diagonal matrix $\mathbf{D}$, where D_{ii} is equal to the number of neighbor of area i, and the proximity matrix $\mathbf{W}$. The proximity matrix $\mathbf{W}$ has non-diagonal elements $w_{ii'} = 1$ for contiguous areas $i \sim i'$. All other elements in $\mathbf{W}$ are zero, and the diagonal matrix $\mathbf{D}$, with diagonal entries d_{ii} equal to the number of areas m_i, that are contiguous to area i. The formulation for the structured spatial effects is known as the intrinsic conditional autoregressive (ICAR) prior [36]. The $\mathbf{v}$ is the vector of unstructured spatial effect, assumed multivariate Normal with mean zero vector, and variance-covariance matrix $\sigma_v^2\mathbf{I^{-1}}$, where σ_v^2 is a variance to be estimated from the data, and $\mathbf{I}$ is $n \times n$ the identity matrix. The $\boldsymbol{\beta}$ is the vector of fixed effects for the covariates effect, assumed multivariate Normal with mean zero vector, and variance-covariance matrix $\sigma_\beta^2\mathbf{I^{-1}}$, where σ_β^2 is a variance to be estimated from the data, and $\mathbf{I}$ is a $K \times K$ identity matrix.
The next model replaces the spatially structured and unstructured effects (u_i and v_i) by the Leroux CAR [37] spatial effect w_i

$$\eta_i = \alpha + w_i + \sum_{k=1}^{k=K} x_{ik}\beta_k \tag{1.11}$$

$$\boldsymbol{w} \sim \text{Normal}(\mathbf{0}, \sigma_w^2\mathbf{R^{-1}}) \tag{1.12}$$

$$\mathbf{R} = \rho_w\mathbf{W} + (1-\rho_w)\mathbf{I}_w$$

where the α and β_k are defined above, and $\boldsymbol{w}$ is the vector of Leroux CAR spatially structured effects, assumed multivariate Normal with mean zero vector, variance parameter σ_w^2, and structure matrix $\mathbf{R}$. The $\mathbf{R}$ matrix is arranged by the spatial dependency parameter ρ_w, the proximity matrix $\mathbf{W}$, and the $n \times n$ identity matrix $\mathbf{I}_w$. The next model replaces the spatially structured effect and the fixed effect for the covariates (w_i and β_k) by the space-varying coefficient β_{ik}

$$\eta_i = \alpha + \sum_{k=1}^{k=K} x_{ik}\beta_{ik} \tag{1.13}$$

$$\boldsymbol{\beta} \sim \text{Normal}(\mathbf{0}, \Sigma_\beta \otimes \mathbf{R}^{-1}) \tag{1.14}$$

In this model, α is already defined, and the variance-covariance matrix is equal to the Kronecker product of a variance covariance matrix Σ_β and the $\mathbf{R}$ matrix defined above.
After presenting the spatial models applied to the dengue data at annual scale, we show the models for spatio-temporal data. We utilized spatio-temporal models of relative risk in Chapter 4, Chapter 5, and Chapter 6. We begin incorporating the time measure $j = 1, \cdots, t$, allowing to extend the SIR for area i and time j

$$SIR_{ij} = \frac{O_{ij}}{E_{ij}} \tag{1.15}$$

where the observed and expected values are now not only in area i but in time j

$$O_{ij} \sim \text{Poisson}(E_{ij} r_{ij}) \tag{1.16}$$

$$r_{ij} = \exp(\eta_{ij}) \tag{1.17}$$

where the Poisson model for the O_{ij} with mean parameter is comprised by the E_{ij} and the relative risk r_{ij}. The first spatio-temporal model explored for our research follows the Knorr-Held taxonomy for spatio-temporal interaction effects models [38]

$$\eta_{ij} = \alpha + u_i + v_i + \gamma_j + \phi_j + \delta_{ij} + \sum_{k=1}^{k=K} x_{ik}\beta_k \tag{1.18}$$

$$\boldsymbol{\gamma} \sim \text{Normal}(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}_\gamma^{-\mathbf{1}}) \tag{1.19}$$

$$\boldsymbol{\phi} \sim \text{Normal}(\mathbf{0}, \sigma_\phi^2 \mathbf{Q}_\phi^{-\mathbf{1}}) \tag{1.20}$$

$$\boldsymbol{\delta} \sim \text{Normal}(\mathbf{0}, \sigma_\delta^2 \mathbf{Q}_\delta^{-\mathbf{1}}) \tag{1.21}$$

where the linear predictor η_{ij} is equal to the additive effect of the overall risk α, the spatially structured effect u_i, and the spatially unstructured effect v_i already defined. In addition to the these parameters, the model includes the $\boldsymbol{\gamma}$ vector of temporally unstructured effects, assumed multivariate Normal with mean zero vector, the variance parameter σ_γ^2, and the $t \times t$ $\mathbf{I}_\gamma$ identity matrix . The model also incorporates the $\boldsymbol{\phi}$ vector of temporally structured effects, assumed multivariate Normal with mean mean zero vector, variance parameter σ_ϕ^2, and the $\mathbf{Q}_\phi$ structure matrix defined by the Random Walk 1 or 2 (RW1 or RW2) structure matrix.

$$(\text{RW1})\mathbf{Q}_\phi = \begin{pmatrix} 1 & -1 & & & \\ -1 & 2 & -1 & & \\ & & \ddots & \ddots & \\ & & -1 & 2 & -1 \\ & & & -1 & 1 \end{pmatrix} \tag{1.22}$$

$$(\text{RW2})\mathbf{Q}_\phi = \begin{pmatrix} 1 & -2 & 1 & & & & \\ -2 & 5 & -4 & 1 & & & \\ 1 & -4 & 6 & -4 & 1 & & \\ & \ddots & \ddots & \ddots & & & \\ & & 1 & -4 & 6 & -4 & 1 \\ & & & 1 & -4 & 5 & -2 \\ & & & & 1 & -2 & 1 \end{pmatrix} \tag{1.23}$$

The model also incorporates the $\boldsymbol{\delta}$ vector of interaction effects, assumed multivariate Normal with zero mean vector, the σ_δ^2 variance parameter, and the $\mathbf{Q}_{\boldsymbol{\delta}}$ structure matrix . Knorr-Held [38] provides a framework for the definition of the $\mathbf{Q}_{\boldsymbol{\delta}}$ structure matrix. The first option for the structure matrix is $\mathbf{Q}_{\boldsymbol{\delta}} = \mathbf{I}_\gamma \otimes \mathbf{I}_w$, where $\mathbf{I}_\gamma$ is already defined, and $\mathbf{I}_\mathbf{w}$ is an $n \times n$ identity matrix, providing the structure matrix for the Type I (unstructured)

interaction effects. The second option for the structure matrix is $\mathbf{Q}_{\boldsymbol{\delta}} = \mathbf{Q}_{\phi} \otimes \mathbf{I}_{w}$, where $\mathbf{Q}_{\phi}$ is a $t \times t$ RW1 or RW structure matrix, and $\mathbf{I}_{w}$ is an $n \times n$ identity matrix, introducing the structure matrix for the Type II (temporal) interaction effects. The third option for the structure matrix is $\mathbf{Q}_{\boldsymbol{\delta}} = \mathbf{Q}_{\boldsymbol{u}} \otimes \mathbf{I}_{\phi}$, where $\mathbf{Q}_{\boldsymbol{u}}$ is an $n \times n$ structure matrix defined by $(\mathbf{D} - \mathbf{W})$, and $\mathbf{I}_{\phi}$ is a $t \times t$ identity matrix, producing the structure matrix for the Type III (spatial) interaction effects. The fourth option for the structure matrix is $\mathbf{Q}_{\boldsymbol{\delta}} = \mathbf{Q}_{\phi} \otimes \mathbf{Q}_{\boldsymbol{u}}$, where $\mathbf{Q}_{\phi}$ and $\mathbf{Q}_{\boldsymbol{u}}$ have been already defined, providing the structure matrix for the Type IV (inseparable) interaction effects. Finally, the spatio-temporal model incorporates $\boldsymbol{\beta}$ parameters for the covariates fixed effects, defined above. The last model to discuss in this overview includes two variations with respect to the previous model

$$\eta_{ij} = \alpha + w_i + \gamma_j + \phi_j + \delta_{ij} + \sum_{k=1}^{k=K} x_{jk}\beta_{jk} \tag{1.24}$$

First, the model replaces the ICAR structured and the unstructured spatial effects (u_i and v_i) by the Leroux CAR structured spatial effect w_i, and second, it replaces the fixed coefficient for the covariates by time-varying coefficients for the covariates.
Instead of a multivariate model for spatio-temporal modeling of risk, we can approach a multivariate model for the joint modeling of many events for a fixed period of time [35]. Then, the observed values in the ith area and the pth disease, $p = 1, \ldots, P$ are assumed to have a Poisson distribution with mean parameter μ_{ip}

$$O_{ip} \sim \text{Poisson}(\mu_{ip})$$
$$\log(\mu_{ip}) = \log(E_{ip}) + \alpha_p + \psi_{ip}$$

where E_{ip} are expected values, α_p are intercepts, and ψ_{ip} are parameters modeling the joint structured or unstructured spatial effect. Several models can be explored for the association between the diseases, and those models will be based on the probability structure of the ψ_{ip} parameters. Multivariate joint models of relative risk for dengue and ZVD are explored in Chapter 7.

1.3.4 Bayesian inference

We followed Gelman et al. [39] for the description of Bayesian inference. Bayesian inference starts creating the joint distribution of the observed data $\mathbf{y}$ and the vector of parameters $\boldsymbol{\theta}$. The joint distribution is obtained as the product of the conditional probability distribution of $\mathbf{y}$ given the $\boldsymbol{\theta}$ and the prior distribution $p(\boldsymbol{\theta})$,

$$p(\mathbf{y}, \boldsymbol{\theta}) = p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}) \tag{1.25}$$

and then the model for Bayesian inference follows

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}, \boldsymbol{\theta})}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \tag{1.26}$$

where $p(\boldsymbol{\theta})$ is the prior specification, $\mathrm{p}(\mathbf{y}|\boldsymbol{\theta})$ is the sampling distribution of the data, and $\mathrm{p}(\boldsymbol{\theta}|\mathbf{y})$ is the posterior distribution, and for the discrete $\boldsymbol{\theta}$

$$p(\mathbf{y}) = \sum_{\boldsymbol{\theta}} p(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) \tag{1.27}$$

and for the continuous $\boldsymbol{\theta}$

$$p(\mathbf{y}) = \int p(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta} \tag{1.28}$$

the sampling distribution is also called the likelihood function, and for a given sample of data, any two probability models $p(\mathbf{y}|\boldsymbol{\theta})$ with the same likelihood function generate the same inference for θ, which is called the *likelihood principle*, and this principle is reasonable within the set of models adopted for a particular analysis. Bayesian inference is exploited in all the chapters with the exception of Chapter 8, where the inference is developed using generalized least squares.

1.3.5 Markov Chain Monte Carlo Methods

For many Bayesian models, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is difficult to obtain analytically. In these cases, a numerical approximation to $p(\boldsymbol{\theta}|\mathbf{y})$ is provided by studying Monte Carlo samples from the posterior density $p(\boldsymbol{\theta}|\mathbf{y})$. The Gibbs sampler [40] is a Monte Carlo method to approximate the posterior distribution. Let a vector of parameters $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$ from a probability distribution $\mathrm{p}(\boldsymbol{\theta})$. Given a starting point $\boldsymbol{\theta}^0 = \left(\theta_1^0, \ldots, \theta_p^0\right)$, the Gibbs sampler generates $\boldsymbol{\theta}^s$ from $\boldsymbol{\theta}^{s-1}$:

$$\begin{aligned} \theta_1^s &\sim p(\theta_1^s|\theta_2^{s-1}, \theta_3^{s-1}, \ldots, \theta_p^{s-1}) \\ \theta_2^s &\sim p(\theta_2^s|\theta_1^{s-1}, \theta_3^{s-1}, \ldots, \theta_p^{s-1}) \\ &\vdots \\ \theta_p^s &\sim p(\theta_p^s|\theta_1^{s-1}, \theta_2^{s-1}, \ldots, \theta_{p-1}^{s-1}) \end{aligned} \tag{1.29}$$

The algorithm generates a sequence of parameters $\boldsymbol{\theta}^1, \ldots, \boldsymbol{\theta}^S$, where $\boldsymbol{\theta}^s$ depends on $\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{s-1}$ only through $\boldsymbol{\theta}^{s-1}$. Then, the sequence possess the Markov property, that is $\boldsymbol{\theta}^s$ is conditionally independent of $\boldsymbol{\theta}^0, \ldots, \boldsymbol{\theta}^{s-2}$ given $\boldsymbol{\theta}^{s-1}$.
The Metropolis-Hastings algorithm [41] [42] is other Monte Carlo algorithm to approximate parameters from the posterior distribution.

1. Generate θ^* from $q_s(\theta^*|\theta^s)$
2. compute the acceptance ratio

$$r = \frac{p_0(\theta^*)}{p_0(\theta^s)} \frac{q_s(\theta^*|\theta^s)}{q_s(\theta^s|\theta^*)} \tag{1.30}$$

3. Sample $u \sim \mathrm{Uniform}(0,1)$

4. If $u < r$ set $\theta^{s+1} = \theta^*$, else set $\theta^{s+1} = \theta^s$

Where $p_0(\cdot)$ is the target distribution, and $q_s(\cdot|\cdot)$ is a proposal distribution. In the first step, a sample θ^* is generated from the proposal distribution. Then, the target distribution is evaluated at the actual and the previous generated value, and the acceptance ratio is calculated. A sequence of parameters is generated with the Markov property.
The slice sampler [43] is also employed to approximate a posterior distribution using a Monte Carlo method. Slice sampling algorithm for univariate distribution follows

1. $u \sim \text{Uniform}(0, p(\theta^s))$.
2. Define a horizontal slice $S = \{\theta : u < p(\theta^s)\}$
3. Define a interval $I = (L, R)$ around θ^s, containing all or much of the slice.
4. Sample θ^{s+1} from Uniform distribution within the I interval.

First step draws a value u^{s+1} from the Uniform distribution $(0, p(\theta^s))$. Then, a new point θ^{s+1} is drawn from an interval $I = (L, U)$ around θ^s that contains all, or much, of the slice within the I interval.
All these MCMC methods can be combined to provide samples from the posterior distribution of the conditional parameters. A Bayesian model can be represented in a directed acyclic graph (DAG) where each model component corresponds to a node and links between nodes display direct dependence [44] [45] . The DAG is *directed* because the relation between two nodes v is defined by an arrow, and *acyclic* because by following the arrows it is not feasible to return to a node after leaving it. A node v is a parent of node w if an arrow starts from v and ends in w, then w is a child of v. The joint distribution of the set of all nodes V in any DAG can be factorized as follows

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto p(\boldsymbol{\theta}, \mathbf{y}) = p(V) = \prod_{v \in V} p(v|parents[v]) \tag{1.31}$$

The full conditional distribution of each node conditioned on the values of the other nodes in the graph is similar to the conditional distribution needed for Gibbs sampling, so Bayesian inference can be implemented using DAGs. The conditional distribution of the node v given the rest of nodes is factorized as the product of the distribution of the node v given their parents, and the distribution of the w given their parents, where the w corresponds with the children of v.

$$p(v|V_{-v}) \propto p(v|parents[v]) \times \prod_{w \in children[v]} p(w|parents[w]) \tag{1.32}$$

The full conditional distribution for any node v is a local computation on the graph, thus only requiring to consider a small part of the model. DAG for Bayesian computations are implemented in the Bayesian software WinBUGS, OpenBUGS [45] and JAGS [46]. MCMC methods are applied in Chapter 2, Chapter 3, Chapter 4, and Chapter 7. A novel Monte Carlo method based on Hamiltonian systems is deferred to its own subsection.

1.3.6 Hamiltonian Monte-Carlo

A novel technique for parameter estimation of Bayesian models is based on Hamiltonian system. Hamiltonian systems are the modern representation of models for classical mechanics. A system of differential equations

$$\frac{dp}{dt} = f(p,q) \tag{1.33}$$

$$\frac{dq}{dt} = g(p,q) \tag{1.34}$$

is a Hamiltonian system if there exists a function $H(p,q)$ such as

$$f(p,q) = \frac{\partial H}{\partial q} \tag{1.35}$$

$$g(p,q) = -\frac{\partial H}{\partial p} \tag{1.36}$$

when such a function $H(p,q)$ exists, we say that $H(p,q)$ is a Hamiltonian function. Then, for approximating parameters for Bayesian models, we can interpret the Hamiltonian function composed by d-dimensional position and momentum vectors (θ and $\mathbf{p}$ respectively), a function $U(\theta)$ denoting the potential energy, and a function K($\mathbf{p}$) denoting the kinetic energy, then a Hamiltonian system is defined by $H(\theta,\mathbf{p}) = U(\theta) + K(\mathbf{p})$. We assumed that θ is the vector of model parameters, the potential energy $U(\theta)$ is proportional to the minus the log probability density of the model parameters, the $\mathbf{p}$ vector are augmented variables, the kinetic energy $K(\mathbf{p})$ is calculated, and then Hamiltonian dynamics are used to update the θ vector in such a way that a Markov chain $\theta_1, \ldots, \theta_N$ is sampled from the target distribution. We follow Hoffman and Gelman [47] to present the Hamiltonian Monte Carlo (HMC) algorithm, where θ^0 is a vector of initial values for the model parameters, ε is the step size, L is the logarithm of the joint density of the variables of interest, S is the number of steps for the Leapfrog algorithm, and N is the iterations number, then the HMC algorithm proceeds:

For $n = 1$ to N do

Sample $\mathbf{p}^0 \sim \text{Normal}(\mathbf{0}, \mathbf{I})$

Set $\theta^m \leftarrow \theta^{m-1}$.

Set $\tilde{\theta} \leftarrow \theta^{m-1}$.

Set $\tilde{\mathbf{p}} \leftarrow \mathbf{p}^0$.

For $i = 1$ to S do

Set $\tilde{\theta}, \tilde{\mathbf{p}} \leftarrow \text{Leapfrog}(\tilde{\theta}, \tilde{\mathbf{p}}, \varepsilon)$

end for

With probability $\alpha = \min\left\{1, \frac{\exp(L(\tilde{\theta}) - \frac{1}{2}\tilde{\mathbf{p}} \cdot \tilde{\mathbf{p}})}{\exp(L(\theta^{m-1}) - \frac{1}{2}\mathbf{p}^0 \cdot \mathbf{p}^0)}\right\}$, set $\theta^m \leftarrow \tilde{\theta}$, $\mathbf{p}^m \leftarrow -\tilde{\mathbf{p}}$

end for

where $\mathbf{I}$ is the identity matrix. For each n iteration, first, momentum variables are re-sampled from a standard multivariate Normal distribution, and then, the Leapfrog algorithm simulates the evolution over time of the Hamiltonian dynamics of the system via the Störmer-Verlet integrator [49]. A number of S Leapfrog steps update the position and momentum variables θ and $\mathbf{p}$ generating a proposal position-momentum pair $\tilde{\theta}$ and $\tilde{\mathbf{p}}$. Then, $\theta^n = \tilde{\theta}$ and $\mathbf{p}^n = \tilde{\mathbf{p}}$ is accepted or rejected following the Metropolis algorithm. The Leapfrog algorithm follows

Function Leapfrog$(\theta, \mathbf{p})$

Set $\tilde{\mathbf{p}} \leftarrow \mathbf{p} + \frac{\varepsilon}{2}\nabla_\theta L(\theta)$.

Set $\tilde{\theta} \leftarrow \theta + \varepsilon\tilde{\mathbf{p}}$.

Set $\tilde{\mathbf{p}} \leftarrow \tilde{\mathbf{p}} + \frac{\varepsilon}{2}\nabla_\theta L(\tilde{\theta})$.

return $\tilde{\theta}, \tilde{\mathbf{p}}$.

The ∇_θ operator is the gradient of the log probability density L with respect to θ. The HMC algorithm requires the need to tune the step size ε and the number of steps S. The No-U-Turn sampler (NUTS) is an extension of the HMC that automatically specify the necessary value of S. HMC algorithm is the base for the Stan software for Bayesian analysis [48]. The HMC approach is employed for estimation of nonlinear parameters in Chapter 9.

1.3.7 Integrated Nested Laplace Approximation

The integrated nested Laplace approximations (INLA) [50] [51] is other numerical method to develop Bayesian analysis. The method is supported in the observation that the data vector $\mathbf{y}$ is conditioned on a linear predictor $\boldsymbol{\eta}$ containing fixed and random effects, and

the linear predictor for many statistical models is a Gaussian latent field with zero mean vector and variance covariance matrix $\mathbf{\Sigma}$ (precision $\mathbf{Q}$).

$$\mathbf{y} \sim p(\mathbf{y}|\boldsymbol{\eta}) = \prod_i p(y_i|\eta_i) \tag{1.37}$$

$$\boldsymbol{\eta} = \mu\mathbf{1} + \mathbf{A}\boldsymbol{\beta} + \sum_i \mathbf{B}_i\mathbf{v}_i + \boldsymbol{\varepsilon} \tag{1.38}$$

The data conditioned by the linear predictor, with intercept μ, covariate matrix $\mathbf{A}$, fixed effects matrix $\boldsymbol{\beta}$, weights matrix $\mathbf{B}_i$, vector of random effects $\mathbf{v}$, and Gaussian noise ε.

$$\mathbf{x} = (\boldsymbol{\mu}, \boldsymbol{\beta}, \mathbf{v}, \boldsymbol{\eta}) \tag{1.39}$$

where $\mathbf{x}$ is a latent Gaussian field. A reformulation of the data model is

$$\mathbf{y}|\mathbf{x}, \boldsymbol{\lambda} \sim \prod_i p(y_i|\mathbf{x}_i, \boldsymbol{\lambda}) \tag{1.40}$$

$$\mathbf{x}|\boldsymbol{\lambda} \sim p(\mathbf{x}|\boldsymbol{\lambda}) = \text{Normal}(\mathbf{0}, \Sigma(\boldsymbol{\lambda})) \tag{1.41}$$

$$\boldsymbol{\lambda} \sim p(\boldsymbol{\lambda}) \tag{1.42}$$

We have the data conditioned on a latent field $\mathbf{x}$ and hyper-parameters $\boldsymbol{\lambda}$. Then, the latent field conditioned on the hyper-parameters is multivariate Gaussian with mean zero, and variance covariance matrix $\Sigma(\boldsymbol{\lambda})$ (precision matrix $\mathbf{Q}(\boldsymbol{\lambda})$). A Bayesian model typically relates the posterior distribution proportional to the hyper-parameters, the latent field conditioned on the hyper-parameters, and the data conditioned on the latent field.

$$p(\mathbf{x}, \boldsymbol{\lambda}|\mathbf{y}) \propto p(\boldsymbol{\lambda}) p(\mathbf{x}|\boldsymbol{\lambda}) \prod_{i \in I} p(y_i|x_i) \tag{1.43}$$

$$\propto p(\boldsymbol{\lambda}) |\mathbf{Q}(\boldsymbol{\lambda})|^{n/2} \exp\left\{ -\frac{1}{2}\mathbf{x}'\mathbf{Q}(\boldsymbol{\lambda})\mathbf{x} + \sum_{i \in I} \log\{y_i|x_i, \boldsymbol{\lambda}\} \right\} \tag{1.44}$$

From the Bayesian model, we are mainly interested in the posterior marginals of the latent Gaussian field and hyperparameters $p(x_i|\mathbf{y})$ and $p(\lambda_i|\mathbf{y})$ respectively.

$$p(x_i|\mathbf{y}) \sim \int p(x_i|\boldsymbol{\lambda}, \mathbf{y}) p(\boldsymbol{\lambda}|\mathbf{y}) d\boldsymbol{\lambda} \tag{1.45}$$

$$p(\lambda_j|\mathbf{y}) = \int p(\boldsymbol{\lambda}|\mathbf{y}) d\boldsymbol{\lambda}_{-j} \tag{1.46}$$

Then for parameter estimation, the INLA approach proceeds by constructing nested approximations to the posteriors marginals

$$\tilde{p}(x_i|\mathbf{y}) \sim \int \tilde{p}(x_i|\boldsymbol{\lambda}, \mathbf{y}) \tilde{p}(\boldsymbol{\lambda}|\mathbf{y}) d\boldsymbol{\lambda} \tag{1.47}$$

$$\tilde{p}(\lambda_j|\mathbf{y}) = \int \tilde{p}(\boldsymbol{\lambda}|\mathbf{y}) d\boldsymbol{\lambda}_{-j} \tag{1.48}$$

where $\tilde{p}(\cdot|\cdot)$ is an approximated conditional density. The marginal posterior of the hyper-parameters $\tilde{p}(\boldsymbol{\lambda}|\mathbf{y})$ is approximated by:

$$\tilde{p}(\boldsymbol{\lambda}|\mathbf{y}) \propto \left.\frac{p(\mathbf{x},\boldsymbol{\lambda},\mathbf{y})}{\tilde{p}_G(\mathbf{x}|\boldsymbol{\lambda},\mathbf{y})}\right|_{\mathbf{x}=\mathbf{x}^*(\boldsymbol{\lambda})} \tag{1.49}$$

where $\tilde{p}_G(\mathbf{x}|\boldsymbol{\lambda},\mathbf{y})$ is the Gaussian approximation to the full conditional $\mathbf{x}$ and $\mathbf{x}^*(\boldsymbol{\lambda})$ is the mode of the full conditional for $\mathbf{x}$, for a given $\boldsymbol{\lambda}$. This is equivalent to the Laplace approximation of the marginal posterior density. Then, the interest is to approximate the posterior conditional distribution of the latent field $\mathbf{x}$. The same Gaussian approximation can be applied to estimate the marginals from $\tilde{p}(\mathbf{x}|\boldsymbol{\lambda}\mathbf{y})$. However, a better approximation is obtained by splitting the vector of parameters as $\mathbf{x} = (x_i, \mathbf{x}_{-i})$ and employing again the Laplace approximation

$$p(x_i|\boldsymbol{\lambda},\mathbf{y}) = \frac{p((x_i,\mathbf{x}_{-i})|\boldsymbol{\lambda},\mathbf{y})}{p(\mathbf{x}_{-i}|x_i,\boldsymbol{\lambda},\mathbf{y})} \approx \left.\frac{p(\mathbf{x},\boldsymbol{\lambda}|\mathbf{y})}{\tilde{p}(\mathbf{x}_{-i}|x_i,\boldsymbol{\lambda},\mathbf{y})}\right|_{\mathbf{x}_{-i}=\mathbf{x}^*_{-i}(x_i,\boldsymbol{\lambda})} \tag{1.50}$$

$$= \tilde{p}(x_i|\boldsymbol{\lambda},\mathbf{y}) \tag{1.51}$$

where $\tilde{p}(\mathbf{x}_{-i}|x_i,\boldsymbol{\lambda},\mathbf{y})$ is the Gaussian approximation of $p(\mathbf{x}_{-i}|x_i,\boldsymbol{\lambda},\mathbf{y})$ and $\mathbf{x}^*_{-i}(x_i,\boldsymbol{\lambda})$ is its mode. Although this strategy works wells, it is costly computationally. A most efficient algorithm is called simplified Laplace approximation, which uses a Taylor series expansion of the Laplace approximation $\tilde{p}(x_i|\boldsymbol{\lambda},\mathbf{y})$. Without regard the estimation strategy (Gaussian, Laplace or simplified Laplace), the integrated nested Laplace approximation proceeds first, by estimating the marginal joint posterior for the hyper-parameters $\tilde{p}(\boldsymbol{\lambda}|\mathbf{y})$, locating the mode, and then performing a grid search to produce a k set of points $\lambda*$, with a corresponding set of weights Δ_k. The marginal posteriors of the hyper-parameters $\tilde{p}(\lambda^*|\mathbf{y})$ are obtained using interpolation and corrected for skewness using log-splines. The conditional posterior $\tilde{p}(x_i|\boldsymbol{\lambda},\mathbf{y})$ is approximated by the Gaussian marginal derived from $\tilde{p}_G(\mathbf{x}|\boldsymbol{\lambda},\mathbf{y})$

$$\tilde{p}(x_i|\boldsymbol{\lambda},\mathbf{y}) = \text{Normal}(x_i;\mu_i(\boldsymbol{\lambda}),\sigma_i^2(\boldsymbol{\lambda})) \tag{1.52}$$

where, $\boldsymbol{\mu}(\boldsymbol{\lambda})$ is the mean vector of Gaussian approximations, and $\sigma_i^2(\boldsymbol{\lambda})$ is a vector of marginal variances. Finally the marginal posterior $\tilde{p}(x_i|\mathbf{y})$ is obtained by numerical integration

$$\tilde{p}(x_i|\mathbf{y}) = \sum_k \tilde{p}(x_i|\boldsymbol{\lambda}_k,\mathbf{y}) \times \tilde{p}(\boldsymbol{\lambda}_k|\mathbf{y}) \times \Delta_k \tag{1.53}$$

The INLA approach is employed for parameter estimation in Chapter 5 and Chapter 6. In Chapter 5, we followed the two-level spatio-temporal modeling based on INLA developed by Ugarte et al [52], while Chapter 5 and Chapter 6 used the parameters constraint specification for parameter identifiability explored by Goicoa et al. [53]

1.3.8 Model selection

For the research articles using Bayesian methods, we employed as selection criteria statistics the deviance information criterion (DIC) [54], the corrected DIC [55], the Watanabe Akaike information criterion (WAIC) [56], the leave-one out (LOO) cross-validation [57], and the logarithmic score (LS) [58]. For the research article using generalized nonlinear least-squares in Chapter 8, we used as selection criteria the Akaike information criterion (AIC) [59] [60]. In addition, model averaging of parameters were accomplished based on the AIC [61].
The DIC is obtained as the expected deviance $\bar{D}$ plus the effective model dimension,

$$\text{DIC} = \bar{D} + d_e \tag{1.54}$$

The deviance is the estimated log predictive density of the data given a point estimate of the fitted model ($\log p(y|\hat{\theta})$) multiplied by -2. The effective number of parameters or model dimension for every data point is equal to

$$d_{ei} = \bar{D}_i - D_i(\bar{\theta}) \tag{1.55}$$

where $D_i(\bar{\theta})$ is the deviance calculated at the mean of the posterior parameters.
It is recognized that the DIC under penalize models containing random effects, so a corrected DIC version is available from Plummer [55].
The log predictive density of the data is unknown because the parameter θ is not known [39]. The predictive accuracy of the fitted model to the data is summarized by the log pointwise predictive density (lpd)

$$\text{lpd} = \sum_{i=1}^{n} \log \int p(y_i|\theta)p(\theta|y)d\theta \tag{1.56}$$

In practice, the lpd is computed by evaluating the expectation using draws from $p(\theta|y)$, the posteriors simulations labeled θ^s, $s = 1,\ldots,S$:

$$\widehat{\text{lpd}} = \sum_{i=1}^{n} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^s) \right) \tag{1.57}$$

The WAIC estimation starts with the $\widehat{\text{lpd}}$ computed log pointwise posterior predictive density and then adding a correction for the effective number of parameters $\hat{p}_{waic}$.

$$\text{WAIC} = \widehat{\text{lpd}} - \hat{p}_{waic} \tag{1.58}$$

$$\hat{p}_{waic} = \sum_{i=1}^{n} V_{s=1}^{S} \log\, p(y_i|\theta s) \tag{1.59}$$

where $V_{s=1}^{S} \log\, p(y_i|\theta s)$ is the posterior variance of the log predictive density for each data point, and $V_{s=1}^{S}$ is the sample variance, $V_{s=1}^{S} a_s = \frac{1}{S-1}\sum_{s=1}^{S}(a_s - \bar{a})^2$. The predictive performance of a model is assessed by generic scoring functions and rules. Measures of

predictive accuracy for point prediction are called scoring functions [62]. The logarithmic score (LS) [37] is a predictive measure calculated based on the conditional predictive ordinates (CPO) method [63]. Letting $\mathbf{y}_{-i}$ correspond to data from all data points except i, then

$$\text{CPO}_i = p(\mathbf{y}_i|\mathbf{y}_{-i}) = \int p(\mathbf{y}_i|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y}_{-i})d\boldsymbol{\theta} \tag{1.60}$$

$$\text{LS} = -\sum_{i=1}^{N} \log(\text{CPO}_i) \tag{1.61}$$

details for the computation of CPO in INLA are available in Held et al. [64], while computation for MCMC are in Lewis et al. [65].
The Leave-one out cross-validation (LOO) can be computed using importance sampling, however, the estimates can be inaccurate. Vehtari et al. [57] proposed a Pareto smoothed importance sampling (PSIS), which allow us to compute LOO using stable importance weights, which is applied in this Thesis.

1.4 Results

In this section, we summarize the main results for the research articles in this Thesis.
In Chapter 2, we explored time varying coefficients models for the disease counts of dengue disease and meteorological variables. We found that the best model included all the covariates with first order random walk time-varying coefficients, and the first order overall trend. In addition, the short-term predictive performance of the selected model was relatively accurate ($< 25\%$ error) predicting weekly dengue case counts at one or two weeks ahead, with the predictions strongly influenced by volatility in the weeks preceding the prediction periods.
In Chapter 3, we found the normalized density vegetation index (NDVI) associated to high dengue risk by census section. The modeling process produced relative risk maps of dengue disease, allowing us to identify areas with high risk in the city. We found slight to fair agreement in high risk by census section between the global aggregated model and the annual aggregated models, and between years in the 2008-2015 period.
In Chapter 4, we discover that the best model was the type IV (inseparable) interaction effects model including constant coefficient for the lag-zero epidemiological week land surface temperature, meaning that the relative risk of dengue disease in every census area depends not only on the relative risk of the neighboring census area, but also on the relative risk from the same area and its neighboring areas in the previous time period.
Chapter 5 reports that for the two-level modeling of relative risk of dengue, the selected best model includes spatially structured random effects for both census sector and commune levels, a temporally structured random effect for the epidemiological periods, and a completely structured interaction term (Type IV) over the census sectors.
In Chapter 6, we provide parallel spatio-temporal risk estimation of ZVD and dengue for the department of Santander and the city of Bucaramanga, using the epidemiological

week as a time measure. We found that at departmental level, the best model included Type II (temporal) spatio-temporal interaction effects, while at city level, the best model included Type IV (inseparable) spatio-temporal interaction effects.
In Chapter 7, we observe that the best joint model at city level was the Generalized Multivariate CAR model, where ZVD high-risk areas are conditioned on dengue high-risk areas, meaning that ZVD risk by area is highly associated with dengue risk by area, where dengue high-risk areas determine the presence of ZVD high-risk areas. At departmental level, the best joint model contains CAR distributed Normal shared components of both diseases and IID Normal spatially unstructured random effects for ZVD and dengue. Thus, the risks of dengue and ZVD alone do not express clustered patterns of high-risk municipalities.
In Chapter 8, we applied non-linear models for real time predictions of the final size and the turning point of the ZVD epidemic in selected Colombian cities using generalized least squares algorithm. We encounter that some non-linear models such as the Richards, the three, four and five-parameter logistic, the Weibull and the sigmoid Emax models are useful to provide real time estimation of the final size of the ZVD epidemic in some Colombian cities, together with parameters obtained from model averaging.
In Chapter 9, we showed Bayesian inference applied to non-linear models for real time predictions of the final size and the turning point of the ZVD epidemic in 10 Colombian cities. Real-time estimation and prediction of the final size of the epidemic were possible near to the 50% to 60% of the total outbreak length for most of the explored nonlinear models. The final size parameters from the three and four parameter logistic models, and the final size weighted parameters were the best predictive parameters based on the generation of small absolute percentage errors during the outbreak period.

1.5 Concluding remarks and further research

In this Thesis we have focused on methods for modeling spatio-temporal areal counts of arboviral diseases such as dengue and ZVD.
In Chapter 2. Dynamic models were used for the analysis of time series of dengue counts. We found a challenge in model interpretation, because, there is not a similar and straightforward interpretation of model parameters as in other time series analytical techniques such as ARIMA or regression models. However, we consider that the results reflect the complex association between the meteorological variables and the incidence of dengue cases, which can not be summarized by models with a small number of parameters. Further research will be directed to testing estimation methods such as INLA and Hamiltonian Monte Carlo as well as implementing dynamic models for data for other cities. We also are interested in extending the univariate setting, to time varying dynamic models for multivariate time series data of arboviral diseases.
Chapter 3 illustrated the spatial modeling of dengue counts at small geographic resolution (from 0.01 to 0.9 km^2) by annual scale including covariates derived from satellite data for year 2008-2015. This study highlight the importance of transforming raw spatial data into relative risk maps of dengue disease for planning and implementation of public health decision making.
In the spatio-temporal setting, each of the spatio-temporal interaction effects models of relative risk originated by Knorr-Held [38] were compared in Chapter 4, Chapter 5, and Chapter 6. For parameter estimation, Chapter 4 used MCMC methods while the other two chapters employed the INLA method. Not only, we tested several geographic aggregation levels but also tested several time aggregation periods and diseases. Both methods displayed accurate results with high computational requirements. These studies advanced the merely spatial setting, generating a background risk to find areas susceptible of further research for factors associated to high risk of dengue or ZVD. In particular for Colombia, we have demonstrated that data with spatial and temporal features generated by the public health surveillance system can be post-processed and used as input for statistical models to gain important information useful for controlling and preventing arboviral diseases. There several lines for future research in the spatio-temporal setting: designing observational studies including data from vectorial ecology per aggregation area, with small granularity in the temporal unit; detecting risk in small geographic aggregation levels, in small regions subjected to control and prevention measures; and, using the models not only for arboviral diseases, but for other diseases in the surveillance system, provided that the assumptions for the data sampling distribution are appropriate from the statistical and biological practice.
In contrast to the spatio-temporal structures, a multivariate joint model of relative risk was approached in Chapter 7. Instead of centering the interest in the risk evolution of dengue and ZVD, we explored the fully joint modeling of relative risk for both diseases at city and departmental level, during the 2015-2016 ZVD outbreak. We discover those areas sharing high risk of dengue and ZVD characterized by a probability close to one. Further work will be developed implementing multivariate joint models not only of dengue and

ZVD, but also for other diseases such as Chikungunya, using the INLA method, or the Hamiltonian Monte Carlo.
Chapter 8 and Chapter 9 analyzed time series of cumulative case counts of ZVD using nonlinear model in order to make real time prediction of epidemiological parameters such as the final size and the turning point of an outbreak. We were able to perform analysis withing the likelihood framework, as well as the Bayesian framework. Nonlinear models are mechanistic, based on specific theoretical conditions about the inherent mechanism generating the data. Further research will be orientated towards the nonlinear mixed effects modeling framework, where multiple areas or towns can be included in a multivariate model for joint parameter estimation and prediction.
We would like to end our conclusion with some opportunities for future research. The statistical models presented in this dissertation can be implemented in real-time surveillance system, in such a way that risk and probability risk estimates of arboviral diseases can be generated as supporting tools for public health teams. For instance, In Colombia, Ocampo et al [66] developed an entomological on line surveillance system integrating data from SIVIGILA, entomological data, and socio economic data, in a spatial and temporal context. However, this system produces information merely based on descriptive statistics. In such context, statistical modeling extensions to real-time surveillance system will enrich the information for controlling and preventing infectious diseases such as dengue, ZVD, and the current threat from many other vector-borne diseases in the Americas.
Finally, we would like to recommend to the public health authorities to give attention to three big issues for implementing statistical modeling similar to the models shown in this dissertation. First, it is necessary to have accurate population data for the aggregation level chosen for modeling. Census data in Colombia is updated every ten years, however, for epidemiological purposes, at small geographic resolution, there is a need for statistical and demographic methods to estimate populations at these small scales. Here, a population register, or the use of multiple health data sources, or big data approximations would be helpful to generate the required data. Second, we did not find data about the vector ecology and biology at small temporal and spatial resolution, which would be necessary to use for an integral approach to control and prevention of arboviral diseases. Third, we also found that the integrity of the notification events reported to the surveillance system relies in the ability of the medical staff to detect and register the event, however, many of the probable cases reported to the system, remain at this status without going to be confirmed by laboratory, creating a gap between the observed and the true situation of the event. We think that this situation needs to be approached by the health system, because the scientific community knows that there are many diseases transmitted by mosquitoes with similar clinical symptoms and signs, and these diseases will arrive to Colombia with a high probability, making the generation of health information very uncertain, unless an accurate characterization of the etiological agent is undertaken.

1.6 References

[1] Pan American Health Organization/World Health Organization, Communicable Diseases and Health Analysis/Health Information and Analysis. PLISA Database. Health Situation in the Americas: Basic Indicators 2017. Washington, D.C., United States of America, 2017.

[2] National Health Institute of Colombia. *Methodology of the routinary surveillance statistical operation.* Health Ministry of Colombia, National Health Institute, Bogotá. 2017. 92 pages. [In spanish]. `https://www.ins.gov.co/Direcciones/Vigilancia/Lineamientosydocumentos/Metodolog%C3%ADa%20%20Sivigila.pdf`

[3] Pollett S, Melendrez MC, Maljkovic B, Duchêne IS, Salje H, Cummings DAT, Jarman RG. Understanding dengue virus evolution to support epidemic surveillance and counter-measure development. *Infection, Genetics and Evolution*. 2018. doi:10.1016/j.meegid.2018.04.032

[4] Salles TS, Sá-Guimarães TE, Lima de Alvarenga ES, Guimarães-Ribeiro V, Ferreira de Meneses MD, de Castro-Salles PF, Rocha dos Santos C, do Amaral Melo AC, Soares MR, Fernandes Ferreira D and Ferreira Moreira M. History, epidemiology and diagnostics of dengue in the American and Brazilian contexts: a review. *Parasites & Vectors*, 2018; **11**:264. https://doi.org/10.1186/s13071-018-2830-8

[5] Villar LA, Rojas DP, Besada-Lombana S, Sarti E. Epidemiological Trends of Dengue Disease in Colombia (2000-2011): A Systematic Review. *PLoS Neglected Tropical Diseases* 2015 **9**(3): e0003499. doi:10.1371/journal.pntd.0003499

[6] Instituto Nacional de Salud, Colombia. Boletín Epidemiológico Semanal, número 52 de 2016, 25 Diciembre - 31 Diciembre. https://www.ins.gov.co/buscador-eventos/Paginas/Vista-Boletin-Epidemilogico.aspx

[7] Plourde A, Block EM. A literature review of Zika virus. *Emerging Infectious Diseases*. 2016; **22**(7), 1185-1192.

[8] Pan American Health Organization/World Health Organization. Zika suspected and confirmed cases reported by countries and territories in the Americas Cumulative cases, 2015-2017. Updated as of 5 January 2017. Washington, D.C.: PAHO/WHO; 2017; Pan American Health Organization. Url: www.paho.org

[9] Pacheco O, Beltrán M, Nelson CA et al. Zika Virus Disease in Colombia - Preliminary Report. *N. Engl. J. Med.* 2016. DOI: 10.1056/NEJMoa1604037

[10] Waller L and Gotway C. *Applied spatial statistics for public health data*. John Wiley & Sons, Inc. 2004.

[11] Racloz V, Ramsey R, Tong S, Hu W. Surveillance of dengue fever virus: A review of epidemiological models and early warning systems. *PLoS Neglected Tropical Diseases*. 2012; **6**(5): 1648.

[12] Louis VR, Phalkey R, Horstick O, Ratanawong P, Wilder-Smith A, Tozan Y, Dambach P. Modeling tools for dengue risk mapping - a systematic review. *International Journal of Health Geographics*. 2014**13**(1): 50.

[13] Naish S, Dale P, Mackenzie JS, McBride J, Mengersen K and Tong S. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC Infectious Diseases* 2014; **14**: 167.

[14] Runge-Ranzinger S, McCall PJ, Kroeger A, Horstick O. Dengue disease surveillance: an updated systematic literature review. *Tropical Medicine & International Health* 2014; **19**(9):1116–1160.

[15] Imai C, Hashizume M. A Systematic Review of Methodology: Time Series Regression Analysis for Environmental Factors and Infectious Diseases. *Tropical Medicine and Health.* 2015; **43**(1): 1- 9.

[16] Wiratsudakul A, Suparit P, Modchang C. Dynamics of Zika virus outbreaks: an overview of mathematical modeling approaches. *PeerJ* 2018 **6**:e4526

[17] Krystosik AR, Curtis A, Buritica P, Ajayakumar J, Squires R, Dávalos D, et al. Community context and sub-neighborhood scale detail to explain dengue, chikungunya and Zika patterns in Cali, Colombia. *PLoS ONE.* 2017, **12**, e0181208.

[18] Riou J, Chiara P, Boelle PY. A comparative analysis of Chikungunya and Zika transmission. *Epidemics.* 2017; **19**: 43-52.

[19] Li X, Liu T, Lin L, Song T, Du X, Lin H, Xiao J, He J, Liu L, Zhu G, Zeng W, Guo L, Cao Z, Ma W, Zhang Y. Application of the analytic hierarchy approach to the risk assessment of Zika virus disease transmission in Guangdong Province, China *BMC Infectious Diseases.* 2017; **17**: 65.

[20] Chien L-C, Lin R-T, Liao Y, Sy FS, Pérez A. Surveillance on the endemic of Zika virus infection by meteorological factors in Colombia: a population-based spatial and temporal study. *BMC Infectious Diseases.* 2018 **18**: 180.

[21] Leta S, Beyene TJ, De Clercq EM, Amenu K, Kraemer MUG, Revie CW. Global risk mapping for major diseases transmitted by Aedes aegypti and Aedes albopictus. *International Journal of Infectious Diseases.* 2018, **67** : 25-35

[22] Shearer FM, Longbottom J, Browne AJ, Pigott DM, Brady OJ, Kraemer MUG, Marinho F, Yactayo S, de Araújo VEM, da Nóbrega AA, Fullman N, Ray SE, Mosser JF, Stanaway JD, Lim SS, Reiner Jr RC, Moyes CL, Hay SI, Golding N. Existing and potential infection risk zones of yellow fever worldwide: a modelling analysis *Lancet*. 2018; **6**: e270-278

[23] Aguiar BS, Lorenz C, Virginio F, Suesdek L, Chiaravalloti-Neto F. Potential risks of Zika and chikungunya outbreaks in Brazil: A modeling study. *International Journal of Infectious Diseases*. 2018; **70**: 20-29.

[24] Martínez-Bello DA, López-Quílez A, Torres Prieto A. Bayesian dynamic modeling of time series of dengue disease case counts. *PLoS Neglected Tropical Diseases*. 2017, **11**(7): e0005696.

[25] Martínez-Bello DA, López-Quílez A, Torres Prieto A. Relative risk estimation of dengue disease at small spatial scale. *International Journal of Health Geographics*. 2017, **16**:31.

[26] Martínez-Bello DA, López-Quílez A, Torres Prieto A. Spatiotemporal modeling of relative risk of dengue disease in Colombia. *Stochastic Environmental Research and Risk Assessment*. 2018, **32**(6): 1587.

[27] Adin A, Martínez-Bello DA, López-Quílez A, Ugarte MD.wo-level resolution of relative risk of dengue disease in a hyperendemic city of Colombia. *PLoS ONE*. 2018, **13**(9): e0203382.

[28] Martínez-Bello DA, López-Quílez A, Torres Prieto A. Spatio-temporal modeling of Zika and dengue infections within Colombia. *International Journal of Environmental Research and Public Health*. 2018, **15**: 137.

[29] Sebrango-Rodríguez CR, Martínez-Bello DA, Sánchez-Valdés L, Thilkarathne PJ, Del Fava E, Vand Der Stuyft P, López-Quílez A, Shkedy Z. Real time parameter estimation of Zika outbreaks using model averaging. *Epidemiology and Infection*. 2017, **145**: 2313–2323.

[30] Gamerman D and Freitas Lopes H. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference.* Second edition. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL 33487-2742. 2006

[31] Pinheiro JC and Bates DM. *Mixed-effects models in S and S-PLUS*. Springer-Verlag New York. 2000

[32] Seber G, Wild C. *Nonlinear regression*, Wiley, New York, 1989.

[33] Lin D, Shkedy Z, Yekutieli D, Amaratunga D, Bijnens L. *Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R*, Springer-Verlag Berlin Heidelberg, 2012.

[34] Lawson A. *Bayesian disease mapping : hierarchical modeling in spatial epidemiology.* Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton, FL. 2009.

[35] Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysisi for Spatial Data, Second Edition.* Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL. 2014.

[36] Besag J and Kooperberg C. On conditional and intrinsic autoregressions. *Biometrika*. 1995; **4**:733-46.

[37] Leroux B, Lei X and Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In M. Halloran and D. Berry (eds), *Statistical Models in Epidemiology, the Environment and Clinical Trials*, 1999, pp. 135–78. Springer-Verlag, New York, NY.

[38] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine.* 2000; **19**:2555-2567.

[39] Gelman A, Carlin J, Stern H, Dunson D, Vehtari A and Rubin D. *Bayesian Data Analysis.* Third edition. Chapman & Hall/CRC, Taylor & Francis Group, Boca Raton, FL. 2013.

[40] Geman, S and Geman D. Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transaction on Pattern Analysis and Machine Intelligence.* 1984; **6**:721–741.

[41] Metropolis N, Rosenbluth AW, Rosenbluth MN, Teller M, Teller E. Equation of State Calculations by Very Fast Computing Machines.*Journal of Chemistry and Physics.* 1953, **21**: 1087.

[42] Hastings W. Monte Carlo sampling methods using Markov chains and their application. *Biometrika*. 1970; **57**:97–109.

[43] Neal R. Slice sampling (with discussion). *Annals of Statistics.* 2003; **31**:705–767.

[44] Lunn DJ, Thomas A, Best N, and Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing.* 2000; **10**: 325-337.

[45] Lunn D, Spiegelhalter D, Thomas A, and Best N. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine.* 2009; **28**: 3049-3067.

[46] Plummer, M. *JAGS Version 4.3.0 user manual.* 2017; 74 p.

[47] Hoffman, MD and Gelman A. The No-U-Turn sampler: adaptively setting path lenghts in Hamiltonian Monte Carlo. *Journal of Machine Learning.* 2014, **15**, 1593-1623.

[48] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker M, Guo J, Li P, and Riddell A. Stan: A probabilistic programming language. *Journal of Statistical Software.* 2017; **76**(1). DOI:10.18637/jss.v076.i01

[49] Hairer, E, Lubich, C, and Wanner, G. Geometric numerical integration illustrated by the Stormer-Verlet method. *Acta Numerica*, 2003.

[50] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2009, **71**, 319–392.

[51] Rue H, Riebler A, Sørbye SH, Illian JB, Simpson DP and Lindgren FK. Bayesian Computing with INLA: A Review. *Annual Review of Statistics and Its Application* 2017, **4** (1), 395-421,

[52] Ugarte MD, Adin A, and Goicoa T. Two-level spatially structured models in spatio-temporal disease mapping. *Statistical Methods in Medical Research.* 2016; **25**(4) 1080–1100.

[53] Goicoa T, Adin A, Ugarte MD, Hodges JS. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment.* 2018, **32**, 749-770.

[54] Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B(Statistical Methodology)* 2002; **64**(4):583–639.

[55] Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics* 2008; **9**: 523–539

[56] Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research.* 2010; **11**: 3571–3594.

[57] Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistical Computing.* 2016; **27**: 1413. `http://link.springer.com/article/10.1007/s11222-016-9696-4`

[58] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 2007; **102**(477):359–378.

[59] Akaike H. *Information theory and an extension of the maximum likelihood principle*, in: B. N. Petrov, F. Csaki (Eds.), Second International Symposium on Information Theory, Akadémiai Kiado, Budapest. 1973, pp. 267–281.

[60] Burnham K, Anderson D. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Researchs* 2004; **33**: 261–304.

[61] Claeskens G, Hjort NL. *Model selection and model averaging*, Cambridge University Press, 2008.

[62] Gelman A, Hwang J, and Vehtari A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing*. 2014, **24**(6):997-1016.

[63] Gelfand AE, Dey DK, Chang H. *Model determination using predictive distributions with implementation via sampling-based methods (with discussion)*. Department of Statistics, Stanford University, Tech. Rep., Stanford, California, 462. 1992.

[64] Held L, Schrödle B, Rue H. Posterior and Cross-validatory Predictive Checks: A Comparison of MCMC and INLA. In: Kneib T, Tutz G. (eds) Statistical Modelling and Regression Structures. Physica-Verlag HD. 2010.

[65] Lewis PO, Xie W, Chen MH, Fan Y, Kuo L. Posterior Predictive Bayesian Phylogenetic Model Selection. Systematic Biology. 2014, **63**(3):309–321.

[66] Ocampo CB, Mina NJ, Echavarría MI, Estrada AL,Alexander N, Ramírez JI, Acuña M, Estupiñán L, Caballero A, Navarro A, Aguirre A. VECTOS, sistema de información para la vigilancia entomolǵica en zonas urbanas *Biomédica* 2017; **37**(Supl.3): 31–81

Chapter 2

Bayesian dynamic modeling of time series of dengue disease case counts

Abstract

The aim of this study is to model the association between weekly time series of dengue case counts and meteorological variables, in a high-incidence city of Colombia, applying Bayesian hierarchical dynamic generalized linear models over the period January 2008 to August 2015. Additionally, we evaluate the model's short-term performance for predicting dengue cases.
The methodology shows dynamic Poisson log link models including constant or time-varying coefficients for the meteorological variables. Calendar effects were modeled using constant or first- or second-order random walk time-varying coefficients. The meteorological variables were modeled using constant coefficients and first-order random walk time-varying coefficients. We applied Markov Chain Monte Carlo simulations for parameter estimation, and deviance information criterion statistic (DIC) for model selection. We assessed the short-term predictive performance of the selected final model, at several time points within the study period using the mean absolute percentage error.
The results showed the best model including first-order random walk time-varying coefficients for calendar trend and first-order random walk time-varying coefficients for the meteorological variables. Besides the computational challenges, interpreting the results implies a complete analysis of the time series of dengue with respect to the parameter estimates of the meteorological effects. We found small values of the mean absolute percentage errors at one or two weeks out-of-sample predictions for most prediction points, associated with low volatility periods in the dengue counts.
We discuss the advantages and limitations of the dynamic Poisson models for studying the association between time series of dengue disease and

meteorological variables. The key conclusion of the study is that dynamic Poisson models account for the dynamic nature of the variables involved in the modeling of time series of dengue disease, producing useful models for decision-making in public health.

2.1 Author Summary

Time series analysis of dengue disease case counts are currently employed to establish associations between dengue disease and environmental, socioeconomic and climatic variables and to predict the evolution of dengue epidemics. Nowadays there is acceptance that climatic factors like environmental temperature, rainfall and relative humidity modify the behavior of the dengue vectors, affecting the transmission of the disease. Thus, in the absence of vector data, climatic factors are commonly used to input transmission models of dengue disease on several temporal and spatial scales.
We applied hierarchical Bayesian dynamic generalized models to dengue diseases case counts in a medium-sized city in Colombia, with constant and time-varying coefficients for calendar trend, and constant and time-varying coefficients for meteorological variables (temperature, rainfall, solar radiation and relative humidity). We selected a final model useful for exploring of the time-varying association between climatic variables and dengue, and the short-term out-of-sample predictions of dengue counts within the study period. We illustrate the modeling process so a data analyst on a multidisciplinary research team could integrate a time series model accounting for the time-varying nature of the data.

2.2 Introduction

Dengue is an arboviral disease caused by a *Flavivirus*, leading to high morbidity in children and adults in tropical countries of Asia and Latin America [1]. There are four genetically distinct but antigenically related (different serotypes) dengue viruses named DEN-1, DEN-2, DEN-3, and DEN-4. All serotypes can cause a spectrum of illness ranging from unapparent or mild fever to the potentially fatal syndrome characterized by hemorrhage, fever and shock syndrome [2]. The infective female *Aedes aegypti* mosquito is the main vector involved in transmiting the viruses causing dengue. The mosquito acquires the virus when it feeds on the blood of an infected human. Several studies show that climate is associated with the mosquito ecology, the infectious agents they carry, and the arboviral transmission of dengue disease [3] [4] [5]. Naish *et al.*(2014) [3] reviewed the studies associating climatic factors and dengue transmission, concluding that higher temperatures affect the rate of larval development, shorten the emergence of adult mosquitoes, increase the biting behavior of mosquitoes, and accelerates virus replication within the mosquitoes. Meanwhile, the combined effect of temperature and relative humidity impact mosquito feeding behavior, vector survival and the probability

to be infected and the ability to transmit dengue.
Epidemiological research on dengue incidence is based on passive surveillance data from case reports [5] [6]. Racloz *et al.* (2012) [5] reviewed early warning modelling in dengue disease, concluding that epidemiological modeling is constrained by limited data sources. Authors encouraged the collection of information at the spatial and temporal level of climatic and socio environmental variables to develop models with stronger predictive capabilities, while Runge-Ranzinger *et al.*(2014) [6] concluded that passive surveillance provides the baseline for outbreak alert, which should be strengthened through the definition of appropriate alert thresholds.
For the time series analysis of dengue case counts associated with meteorological variables, diverse methodologies have been employed, including auto-regressive integrated moving average (ARIMA) models [7] [8] [9] [10] [11] [12] [13] [14] [15], Poisson multivariate regression forecasting models [16] [17] [18], distributed lag non-linear models (DLNM) [19] [20], decision trees with cross-validation [21], multi-resolution analysis and fuzzy systems [22], stepwise negative binomial multivariate linear regression analysis [23], wavelet time series analysis [24], probabilistic random walks [25] [26], and dynamic generalized linear models (DGLM) [27] [28] [29].
DGLMs are extensions of the dynamic linear models [30] [31], based on two sets of equations, a measurement or observation equation and the transition or state equations. The observation equation establishes a link between observations and unobserved variables, and the transition equations describe the evolution of state variables. DGLMs allow the inclusion of components modeling seasonality, trend, cyclicity and covariates [31]. The classic models for calendar trend are the first-order random walk model, the local linear trend model (first-order random walk plus trend) and the second-order random walk [32]. Modeling seasonality and cyclicity is accomplished through dummy variables or trigonometric series defined in the transition equations, and covariates are included with constant or time-varying coefficients [32].
DGLM parameter estimations have followed different approaches. Linear Bayes estimation with conjugate updating [30] [31] or iteratively weighted Kalman filter and smoother, accompanied by the expectation-maximization (EM) algorithm for the estimation of unknown hyperparameters [32], was applied by Chiogna and Gaetan [33] to explore the association between pollution covariates and respiratory diseases. Shepard *et al.* [34] applied likelihood base inference for non-Gaussian state space parameters, based on importance sampling.
DGLMs estimated by Markov Chain Monte Carlo (MCMC) simulations have been explored by Gamerman [35], Ferreira and Gamerman [27] (modeling Dengue disease and meningitis with covariates and seasonal terms), Schmidt and Pereira [28] and Alves *et al.* [36] including covariates with constant coefficients for time accompanied by covariates modeled by transfer functions. Malhão *et al.* [29] implemented DGLM for time series of dengue cases, capturing temporal dependencies not explained by covariates, and modeling dengue over-mortality.
Colombia is one of the countries with the highest incidence of dengue disease in the tropics, and it is testing dengue control by vaccination [37], a topic of interest among the research community [38]. The country possesses climatic, environmental and socio-

geographic conditions favoring the growth and development of the dengue vector. The *Aedes aegypti* mosquito is found across more than 80% of the territory, which has an altitude of 1000 m and 2200 m above sea level, and the *Aedes albopictus* (forest and urban dengue vector) has also been reported [39].
Bucaramanga is among the Colombian cities with the highest annual dengue incidence for the 2008-2015 period. In 2010 and 2012 the city experienced incidence rates of 1515 and 279.93 cases per 100,000 people, respectively, while for the same years the incidence rates for the country were 657 and 221.9 cases per 100,000, respectively [39] [40]. The *Aedes aegypti* mosquito has been reported as the dengue vector in the city of Bucaramanga. While vectorial surveillance studies did not exist in 2008-2015 to quantify the presence of vectors, their abundance, occurrence, distribution and other epidemiological parameters at monthly or weekly temporal scales for Bucaramanga, information of climatic variables such as environmental temperature, rainfall, solar radiation, and relative humidity are available from several sources at these temporal scales. These data offer opportunities to analyze the relation between time series of dengue cases and climatic variables, as Rúa-Uribe *et al.*(2013) [8] show for another Colombian city.
The aim of this study is to model the association between time series of dengue case counts and meteorological variables, in a high-incidence city of Colombia, applying Bayesian hierarchical dynamic generalized linear models, during the period January 2008 to August 2015. Additionally, we evaluate the model's performance in short-term prediction of dengue cases.

2.3 Materials and Methods

2.3.1 Data

Bucaramanga is a medium-sized city in Colombia, at 959 meters above sea level, with a population of 527,913 people (projected population, 2015), at the coordinates $7^o07'07"$N, $73^o06'58"$W. We collected dengue case counts for 2008-2015 in metropolitan Bucaramanga from the Surveillance National System of Public Health (SIVIGILA). The total dengue case counts (probable and confirmed cases of dengue and severe dengue plus dengue mortality) by epidemiological week (EW) were computed in the interval between the first EW of January 2008 to the last EW of August 2015, for a total of 396 EW. For the meteorological variables (MV), daily maximum temperature (°C), daily total rain fall (mm), daily maximum solar radiation (Watts/m^2) and daily maximum relative humidity (%) were obtained from three stations of the Defense Corporation of the Bucaramanga Plateau (CDMB). Daily maximum temperature (°C) and daily total rain fall (mm/m^2) were obtained from the Institute of Hydrology, Meteorology and Environmental Studies of Colombia (IDEAM) for two meteorological stations. Daily values for every variable were averaged by EW and by station, and then the weekly averages of all stations were averaged, obtaining one value per MV and EW.

2.3.2 Hierarchical dynamic Poisson models

We fitted Bayesian hierarchical dynamic Poisson models to dengue case counts. Let y_t be the case count for dengue in EW t ($t = 1, \cdots, T$ and $T = 396$), and

$$y_t \sim \text{Poisson}(\lambda_t) \tag{2.1}$$

The logarithm of the mean λ_t is modeled with two options. The first option is the inclusion of a constant coefficient α for the calendar trend,

$$\log(\lambda_t) = \begin{cases} \alpha \\ \alpha + \sum_{j=1}^{J} \beta_j x_{t-1,j} \\ \alpha + \sum_{j=1}^{J} b_{t,j} x_{t-1,j} \end{cases} \tag{2.2}$$

where α is Normal with mean 0 and variance 10, which allows flexibility for the exploration of the parameter space. The second option is the inclusion of time-varying coefficients α_t for the calendar trend,

$$\log(\lambda_t) = \begin{cases} \alpha_t \\ \alpha_t + \sum_{j=1}^{J} \beta_j x_{t-1,j} \\ \alpha_t + \sum_{j=1}^{J} b_{t,j} x_{t-1,j} \end{cases} \tag{2.3}$$

where the time-varying coefficients α_t are defined with Normal random walk 1 (RW1) or Normal random walk 2 (RW2) priors. The Normal RW1 priors for α_t are defined as

$$\begin{aligned} \alpha_1 &\sim \text{Normal}(3, 0.2) \\ \alpha_t &\sim \text{Normal}(\alpha_{t-1}, \tau_\alpha); \quad (2 \leq t \leq T) \end{aligned}$$

and the Normal RW2 priors for α_t follow

$$\begin{aligned} \alpha_1, \alpha_2 &\sim \text{Normal}(3, 0.2) \\ \alpha_t &\sim \text{Normal}(2\alpha_{t-1} - \alpha_{t-2}, \tau_\alpha); \quad (3 \leq t \leq T) \end{aligned}$$

where for the Normal(3,0.2) prior, the mean of 3 for α_1 and α_2 in the exponential scale is close to the observed dengue case counts at time points 1 and 2, and 0.2 is a precision (variance of 20) that allows flexibility for these parameters. τ_α is the precision parameter with Gamma(1,0.1) hyperprior, which represents a Gamma prior noninformative distribution centered at 10 with variance of 100. In Equation 2.2 and Equation 2.3, the $x_{t-1,j}$ ($j = 1, \cdots, J$ and $J = 4$) are the mean centered MVs temperature ($j = 1$), rainfall ($j = 2$), solar radiation ($j = 3$) and relative humidity ($j = 4$). The β_j are constant coefficients for lag-one MV, and $b_{t,j}$ are time-varying coefficients for lag-one MV. Normal priors with mean 0 and variance 10 were assigned to the constant coefficients β for the covariates. The time-varying coefficients for the lag-one covariates received first-order Normal RW1 priors,

$$\begin{aligned} b_{1,j} &\sim \text{Normal}(0, 0.1) \\ b_{t,j} &\sim \text{Normal}(b_{t-1,j}, \tau_{b_j}) \quad (2 \leq t \leq T) \end{aligned}$$

where for the Normal(0,0.1), we let $b_{1,j}$ start centered at zero, with a 0.1 precision (variance of 10), allowing a large space for exploring the parameter. Gamma(1,0.001) prior distributions (Gamma centered at 1000 with variance of 100,000) are assigned to the precision parameters τ_{b_j}. The reason for this prior is that we constrain the variance of the $b_{t,j}$ to be very small, smoothing the trend of the time-varying coefficients and allowing us to visualize the smoothed trend of the covariate effects.

We modeled missing data in the covariates by imputing the empty values, assuming a Normal(μ_{t-1}, τ_j) prior for $t = 1, \cdots, T$ and $T = 396$, where μ_{t-1} is the value of the lag-one week meteorological centered variable, where τ_j is a precision parameter with Gamma(0.1,0.1) priors for temperature, for rainfall, solar radiation, and relative humidity, where the Gamma prior is an informative prior centered at 0.1 with dispersion 10, slightly constraining the imputed values of the covariates to have a small variance, without restricting to high variance values.

Models were fitted applying MCMC using WinBUGS 1.4 software [41], with 3 chains, 50,000 iterations total, 46,000 iterations burn-in and thinning of 4, obtaining a final sample of 1000 iterations per chain. Convergence was assessed by Gelman-Rubin diagnostic [42] and visual inspection of the simulations chains. Model selection was accomplished using deviance information criteria (DIC) [43]. When DIC measures are used for model selection, models with small deviance $\bar{D}$, a small number of parameters p_D and a small DIC are selected for inference.

After fitting all models, and selecting the final model for inferences, we were interested in evaluating the short-term prediction performance of the selected final model.

We obtained predictions at several time points, during the study period T=396. We selected estimation periods 1 to t, where t was in increments of 20 EWs, starting in the 20th EW of the study period and ending in the 380th EW. We obtained 19 upper bounds for the estimation period 1 to t.

Then we fitted models for periods 1 to p, where $p = t + k$ ($k = 1, \cdots, 4$), and the k are prediction periods (one, two, three or four weeks ahead). We used the same conditions defined above for the MCMC simulations. Samples from the posterior predicted distribution for the prediction periods k were obtained, and the mean and 95% credible intervals (CIs) for the cases of dengue were calculated. To evaluate the prediction performance from the final model, we calculated the mean absolute percentage error (MAPE) per MCMC iteration between the predicted cases of dengue y_{pred_k} and the observed case count y_k, at prediction periods k ($\sum_k |(y_{pred_k} - y_k)/y_k|/k$). We present the median MAPE of the posterior predictive distribution for all the estimation periods t for one, two, three and four weeks ahead as a measure of short-term model performance for predicting dengue case counts.

2.4 Results

2.4.1 Exploratory data analysis

The total number of cases of dengue disease for the study period was 26,755. The weekly case count averaged 67.6, with a median of 52 (range 7 to 247). There were three dengue disease outbreaks in 2010, 2013 and 2014, with small case counts in 2011 and 2012 (Figure 2.1). The partial autocorrelation function for the time series of dengue case counts (Figure 2.1) suggest a first- or second-order autoregressive process. Maximum weekly

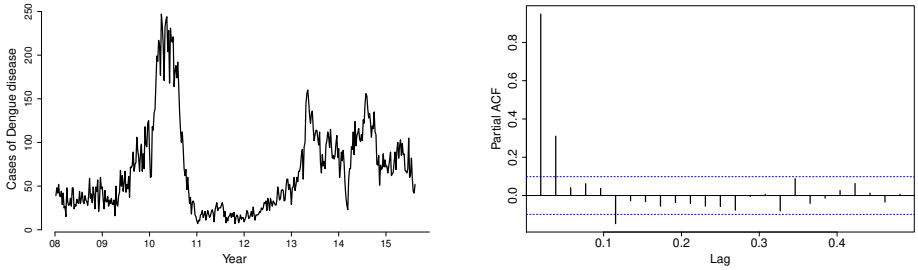


Figure 2.1: Dengue time series plots. Time series plot of dengue case counts (left) and partial autocorrelation function plot of dengue case counts (right).

temperature averaged 27°C, with a minimum of 23.6°C, a maximum of 30.4°C, and 18 missing values. Mean and median values of weekly rainfall were 2.7 mm/m^2 and 3.6 mm/m^2, respectively, with a minimum of 0, a maximum of 24.8 mm/m^2, and 11 missing values. Weekly maximum solar radiation averaged 946.5 Watts/m^2, with median of 940.9 Watts/m^2, a minimum of 733.5 Watts/m^2, a maximum of 1279 Watts/m^2, and 66 missing values. Maximum weekly relative humidity averaged 94.2%, with a minimum of 79.2%, a maximum of 99.5%, and 63 missing values.

Figure 2.2 shows plots of time series for MVs, and plots of the average dengue case counts by intervals of the MVs. While time series for temperature and relative humidity display an upward trend over the 396 EWs, solar radiation decreases, and precipitation shows highly volatile behavior. Dengue disease case counts are positively correlated with temperature, and negatively correlated with solar radiation. There is no apparent association between dengue case counts and precipitation or relative humidity.

In Figure 2.3, linear correlations between the meteorological variables and dengue case counts show positive and moderate correlation with temperature and negative and moderate linear correlation with relative humidity, solar radiation and rainfall. Relative humidity and solar radiation display high positive correlations with their own lag-1 and lag-2 values, followed by temperature and rainfall. Rainfall, relative humidity and solar radiation are positively and moderately correlated, while rainfall and temperature show negative and moderate correlation. Finally, we highlight the negative and low correlation between solar radiation and temperature.

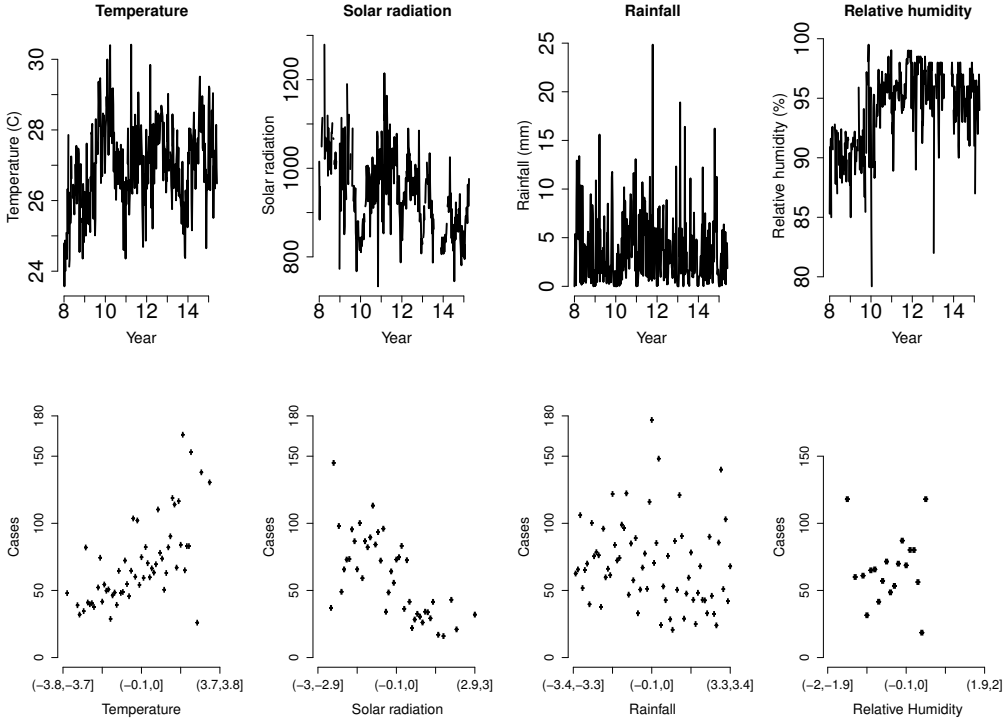


Figure 2.2: Meteorological variables time series plots. Time series plots of temperature, rainfall, solar radiation and relative humidity (top) and scatter plots of the average number of cases of dengue by intervals of the meteorological variables (bottom)

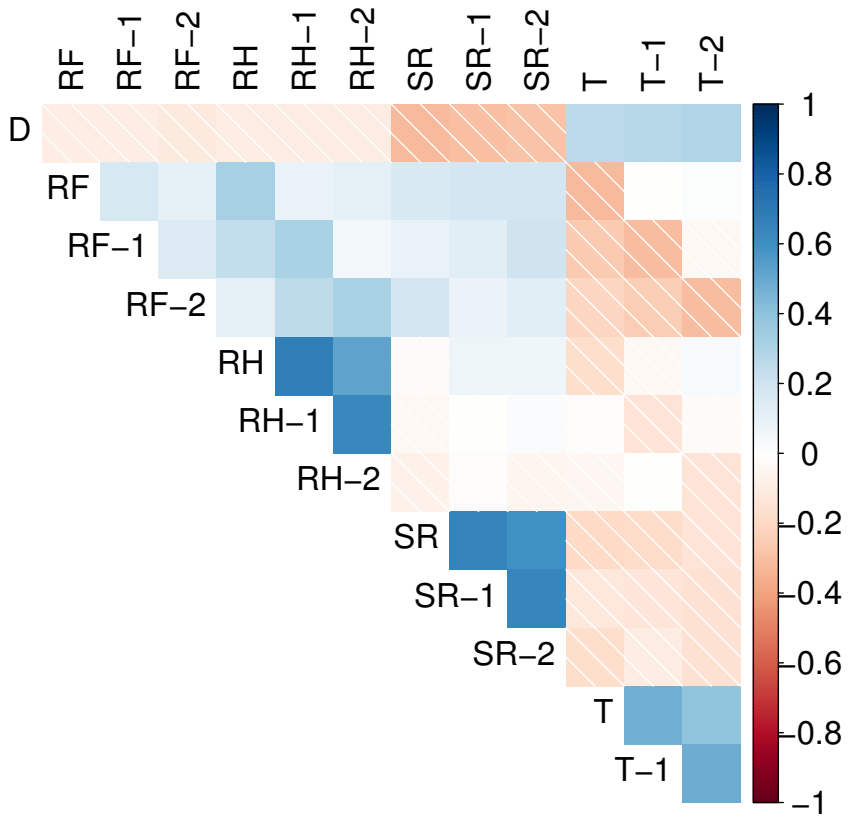


Figure 2.3: Correlation matrix plot of weekly dengue case counts and lag-zero, lag-one and lag-two meteorological variables. D: dengue disease cases. RF: rainfall. RH: relative humidity. SR: solar radiation. T: temperature.

2.4.2 Dynamic Poisson models

In this section, we begin by presenting the results from the models without covariates (only constant coefficient (CC) (α) or RW1 or RW2 time-varying coefficients (TVCs) (α_t) for calendar trend). We define calendar trend as the pattern observed in the model's parameters over the EWs in the entire study period (2008-2015), not the trends observed over any given epidemiological year. We then present the results from models including CC (β_j) for covariates, and CC (α) or RW1 or RW2 TVCs (α_t) for calendar trend. Finally, we exhibit the results from models including RW1 TVCs ($b_{t,j}$) for the covariates with CC (α) or RW1 or RW2 TVCs (α_t) for calendar trend.

Models without covariates

For the models without covariates, the deviance and DIC for the model with CC (α) for calendar trend are 15,959.8 and 15,960.8, respectively. For the models with RW1 or RW2 TVCs (α_t) for trend, the respective deviance and DIC are 2716.4 and 2901.1 for the RW1 model, and 2901.5 and 2990.0 for the RW2 model. We conclude that the model with CC (α) for trend shows worse fit than the models with RW1 or RW2 TVCs (α) for trend of calendar time. The models with RW1 or RW2 TVCs (α_t) for calendar trend have similar DIC, while the model with RW1 TVCs (α_t) for calendar trend offers the best fit (small deviance).

Models with CC (β_j) for the covariates

Table 2.1 presents the DIC selection measures from the simple (single covariate) Poisson regression models with CC (β_j) for the covariates, and CC (α) or RW1 or RW2 TVCs (α_t) for calendar trend.

First, for every meteorological variable, the model with CC (α) for calendar trend and CC (β_j) for the covariates corresponds to the simple Poisson regression, while the models with RW1 or RW2 TVCs (α_t) for trend and CC (β_j) for the covariates are the simple dynamic Poisson regression.

Second, the simple Poisson regression models display worse fit than the simple dynamic Poisson regression models, evidenced by high DIC and deviance values.

Third, the fit of the simple Dynamic Poisson models with CC (β_j) for the covariates, and RW1 TVCs (α_t) for calendar trend is better than models with RW2 TVCs (α_t) for calendar trend.

Table 2.2 displays parameter estimates of the CC (β_j) for the covariates, from models with CC (α) or RW1 or RW2 TVCs (α_t) for calendar trend, from Table 2.1. Parameter estimates for the CC for temperature are 0.207 (95% CI: 0.197, 0.217); solar radiation, -0.309 (95% CI: -0.324, -0.294); and rainfall -0.026 (95% CI: -0.030, -0.022), from models with CC (α) for calendar trend suggesting a strong association between these

Table 2.1: DIC measures for models with constant coefficient (α), RW1 or RW2 TVCs (α_t) for calendar trend with CC (β_j) for the covariates

	$\beta_{\text{Temperature}}$			β_{Rainfall}		
Trend	$\bar{D}$	p_D	DIC	$\bar{D}$	p_D	DIC
α	14341.9	12.5	14354.4	15735.2	6.3	15741.4
α_t (RW1)	2713.4	188.9	2902.3	2717.4	185.2	2902.6
α_t (RW2)	2836.2	119.5	2955.8	2908.3	88.7	2997.0
	$\beta_{\text{Solar radiation}}$			$\beta_{\text{Relative humidity}}$		
	$\bar{D}$	p_D	DIC	$\bar{D}$	p_D	DIC
α	13627.1	46.3	13673.4	15260.3	-590.4	14669.9
α_t (RW1)	2717.4	184.9	2902.3	2717.7	184.9	2902.7
α_t (RW2)	2905.6	89.5	2995.1	2898.9	90.8	2989.8

Table 2.2: Parameter estimates of models with CC (α) or RW1 or RW2 TVCs (α_t) for calendar trend and CC (β_j) for the covariates.1

Trend	Mean	SD	95% CI	Trend	Mean	SD	95% CI
	$\beta_{\text{Temperature}}$				$\beta_{\text{Solar radiation}}$		
α	0.207	0.005	(0.197, 0.217)	α	-0.309	0.007	(-0.324, -0.294)
α_t (RW1)	0.010	0.013	(-0.014, 0.035)	α_t (RW1)	-0.001	0.022	(-0.046, 0.040)
α_t (RW2)	0.006	0.011	(-0.017, 0.027)	α_t (RW2)	-0.010	0.022	(-0.047, 0.035)
	β_{Rainfall}				$\beta_{\text{Relative humidity}}$		
α	-0.026	0.002	(-0.030, -0.022)	α	0.009	0.026	(-0.029, 0.031)
α_t (RW1)	0.001	0.003	(-0.005, 0.007)	α_t (RW1)	-0.003	0.005	(-0.012, 0.007)
α_t (RW2)	0.001	0.002	(-0.004, 0.006)	α_t (RW2)	-0.007	0.004	(-0.015, 0.000)

variables and the weekly case counts of dengue.
There is no statistical association between cases of dengue disease and relative humidity (0.026, 95% CI: -0.029, 0.031). These parameters correspond to the simple Poisson regression model.
Although models with CC (α) for calendar trend show strong statistical association between covariates and dengue, the point estimates and 95% CIs from models with RW1 or RW2 TVCs (α_t) for trend show a weak association between cases of dengue and the meteorological variables, while these models present the best fit (small DIC and deviance).

Models with RW1 TVCs ($b_{t,j}$) for the covariates

Next, we fitted models with CC (α) or RW1 or RW2 TVCs (α_t) for calendar trend, with RW1 TVCs ($b_{t,j}$) for the lag-one covariates. Information criteria for these simple dynamic Poisson regression models with TVCs ($b_{t,j}$) for the covariates are presented in Table 2.3.
For temperature, DIC for the models with CC (α) or RW2 TVCs (α_t) for calendar trend are higher than the model with RW1 TVCs (α_t) for calendar trend.

Table 2.3: DIC measures for models with CC (α) or RW1 or RW2 TVCs (α_t) for calendar trend with RW1 TVCs ($b_{t,j}$) for the covariates.

	$b_{t,}$Temperature			$b_{t,}$Rainfall		
Trend	$\bar{D}$	p_D	DIC	$\bar{D}$	p_D	DIC
α	2872.0	314.8	3186.8	2989.4	322.0	3311.5
α_t (RW1)	2710.2	187.7	2897.9	2706.7	189.4	2896.1
α_t (RW2)	2841.3	103.5	2944.8	2846.2	114.3	2960.5
	$b_{t,}$Solar radiation			$b_{t,}$Relative humidity		
Trend	$\bar{D}$	p_D	DIC	$\bar{D}$	p_D	DIC
α	3177.6	38.8	3216.4	3030.0	-539.1	2490.9
α_t (RW1)	2709.7	172.4	2882.1	2705.7	176.9	2882.5
α_t (RW2)	2783.8	84.6	2868.4	2807.6	94.2	2901.8

DIC for rain fall display similar results as temperature, i.e., DIC for the model with CC (α) or RW2 TVCs (α_t) for trend are higher than the model with RW1 TVCs (α_t) for calendar trend.
For solar radiation, DIC for the model with RW2 TVCs (α_t) for calendar trend is smaller than the models with RW1 TVCs (α_t) and CC (α) for calendar trend.
Lastly, the model with RW1 TVCs for relative humidity plus CC (α) for calendar trend have the smallest DIC for this covariate (DIC = 2490.9), but the number of parameters (p_D) is negative (p_D = -539.1), which makes this model a poor option. DIC from the models with RW1 or RW2 TVCs (α_t) for calendar trend do not present negative p_D. The smallest DIC is for the model with RW1 TVCs (α_t) for calendar trend.
At this stage of the analysis, we identified models with RW1 TVCs (α_t) for calendar trend plus RW1 TVCs ($b_{t,j}$) for the covariates, as the models offering the best fit (smallest deviance and DIC). Then, in addition to the simple dynamic Poisson regression models with TVCs ($b_{t,j}$) for the covariates, we fitted multiple (multiple variables) dynamic Poisson models, presenting the information criteria in Table 2.4. DIC measures for all the models with RW1 TVCs (α_t) for trend plus RW1 TVCs ($b_{t,j}$) for the meteorological variables range from 2831.4 to 2897.6 (Table 2.3). The model with RW1 TVCs for solar radiation and relative humidity ($b_{t,SR} + b_{t,RH}$) presents the smallest DIC (DIC = 2831.4) and effective number of parameters (p_D = 133.5), followed by the model including all the MVs in the predictors ($b_{t,T} + b_{t,RF} + b_{t,SR} + b_{t,RH}$) (DIC = 2847.2), which presents the smallest deviance, selecting this saturated model for inference instead of model with solar radiation and relative humidity, because the model with the lowest DIC is also the model with the most imputed variables (solar radiation and relative humidity).
We include convergence diagnostic measures in the Supplementary appendix for the model parameters in Table 2.4.
Finally, from the model with TVCs for all the meteorological variables ($b_{t,T} + b_{t,RF} + b_{t,SR} + b_{t,RH}$) in Table 2.3, we plot the time-varying parameter estimates (mean and 95% CIs) in Figure 2.4.
TVCs for temperature and solar radiation present higher variability than the coefficients for relative humidity and rainfall. Point estimates for temperature start at values higher

Table 2.4: DIC selection measures from models with RW1 TVCs (α_t) for calendar trend and RW1 TVCs ($b_{t,j}$) for the covariates. $b_{t,T}$: temperature. $b_{t,RF}$: rainfall. $b_{t,SR}$: solar radiation. $b_{t,RH}$: relative humidity.

Model	$\bar{D}$	p_D	DIC
$b_{t,T}$	2710.2	187.7	2897.9
$b_{t,RF}$	2706.7	189.4	2896.1
$b_{t,SR}$	2709.7	172.4	2882.1
$b_{t,RH}$	2705.7	176.9	2882.5
$b_{t,T} + b_{t,RF}$	2699.7	196.0	2895.7
$b_{t,T} + b_{t,SR}$	2701.8	183.3	2885.1
$b_{t,T} + b_{t,RH}$	2699.8	179.4	2879.2
$b_{t,P} + b_{t,SR}$	2696.9	182.8	2879.7
$b_{t,RF} + b_{t,RH}$	2694.8	182.8	2877.6
$b_{t,SR} + b_{t,RH}$	2697.9	133.5	2831.4
$b_{t,T} + b_{t,RF} + b_{t,SR}$	2687.2	191.3	2878.5
$b_{t,T} + b_{t,RF} + b_{t,RH}$	2686.9	192.5	2879.4
$b_{t,RF} + b_{t,SR} + b_{t,RH}$	2684.8	182.7	2867.6
$b_{t,T} + b_{t,SR} + b_{t,RH}$	2692.8	178.9	2871.7
$b_{t,T} + b_{t,RF} + b_{t,SR} + b_{t,RH}$	2680.9	166.3	2847.2

than zero, in contrast with relative humidity, solar radiation and rainfall, which begin almost at zero. TVCs for temperature are above zero for 2008 and 2010, below zero for 2009 and 2014, and close to zero for 2011 to 2013 and for the year 2015, with 95% CIs not including zero only for 2008.
TVCs for solar radiation are above zero for 2009 and 2015, with a small peak in 2010, and below zero for 2011 to 2014, with the 95% CIs including zero for the entire study period, with the exception of 2009.
For rainfall, TVCs present high volatility, with coefficients above zero for 2009, 2010, 2011, 2014 and 2015, and below zero for 2008, 2012 and 2014, with 95% CIs including zero for all years in the study period except 2009.
TVCs for relative humidity are above zero for 2008, 2009 and 2012 and below zero for 2010, 2011 and 2013; the 95% CIs cross zero for the complete study period.

Short-term prediction of dengue case counts

We use the model with RW1 TVCs (α_t) for calendar trend plus TVCs ($b_{t,j}$) for the covariates ($\log(\lambda_t) = \alpha_t + \sum_{j=1}^{4} b_{t,j}$) ($j$ = 1, temperature; j = 2, rainfall; j = 3, solar radiation; j = 4, relative humidity) to obtain a forecast for several time points during the study period 1 to T (T = 396). Figure 2.5 presents the observed and predicted dengue case counts obtained for the selected final model.
Based on Figure 2.5, we can distinguish the trend of the dengue case counts in the time periods close to the prediction points: from June to December 2008, the trend was stable.

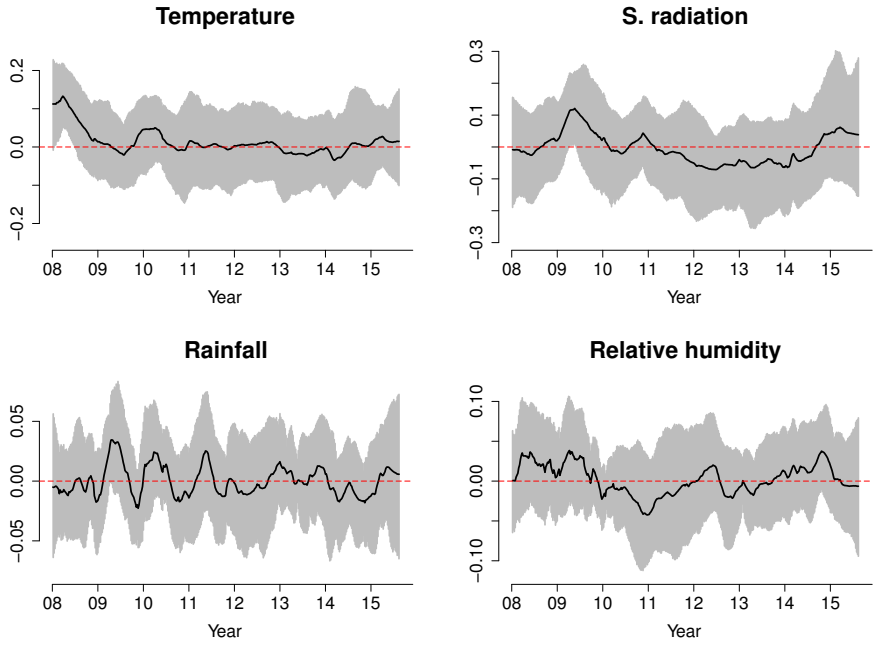


Figure 2.4: Posterior mean and 95% CI for the TVCs ($b_{t,j}$) for temperature, rainfall, solar radiation and relative humidity from the saturated model.

Then, there was a gradual increase in May 2009 and a sharp rise in November 2009. Afterwards, the trend stabilized, but then became highly volatile in May 2010 (at the peak of the 2010 outbreak) before slowly decreasing in November 2010. Between May and October 2011, the trend was stable, showing a slow increase from April to October 2012. The trend from March to September 2013 is a rapid decrease, followed by a rapid increase in March 2014, and a slow decrease in September 2014, before evening out in March 2015.

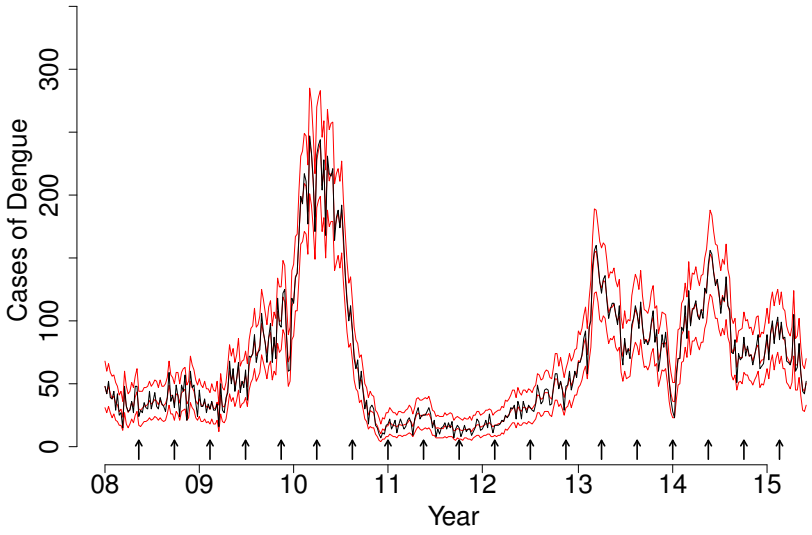


Figure 2.5: Posterior mean and 95% CI for the predicted case counts of dengue disease (red lines) from the selected model, and observed counts (gray line). Arrows representing the EW were short-term predictions of dengue case counts at one, two, three and four weeks.

Table 2.5 presents the MAPE between the predicted mean and the observed dengue case counts for short-term prediction periods at one, two, three and four weeks, estimated at selected EW after the first EW of 2008, from the model selected for inferences. A quick inspection reveals that the highest MAPEs correspond to the EW associated with outbreaks in 2010, 2013 and 2014.

Table 2.5: Posterior median of the MCMC simulations for the mean absolute percentage error (MAPE) to evaluate the short-term predictive performance of the final model in selected EWs after the first EW of January 2008

			Weeks ahead			
Year	Date	EW after first EW 2008	1	2	3	4
2008	May 11	20	12.0	15.0	18.3	20.2
	Sep 28	40	13.0	17.0	18.0	20.5
2009	Feb 15	60	8.0	11.0	12.0	14.0
	Jul 05	80	14.0	16.0	22.3	26.0
	Nov 22	100	26.0	36.0	43.7	43.8
2010	Apr 11	120	50.0	66.0	78.7	84.2
	Aug 29	140	23.0	30.0	39.3	42.0
2011	Jan 01	160	5.0	7.5	8.0	8.2
	Jun 06	180	7.0	8.0	10.0	10.8
	Oct 10	200	4.0	6.5	6.0	6.8
2012	Mar 03	220	4.0	5.5	6.0	6.8
	Jul 07	240	6.0	8.0	9.0	10.2
	Dec 12	260	15.0	16.5	20.0	20.0
2013	Jun 05	280	30.0	33.5	36.0	41.2
	Sep 09	300	20.0	26.0	30.0	34.2
2014	Feb 02	320	17.0	20.5	20.3	23.0
	Jun 06	340	33.0	39.0	43.3	45.8
	Nov 11	360	17.0	19.5	23.7	24.2
2015	Apr 04	380	19.0	23.0	25.7	26.8

Figure 2.6 show the MAPE results presented in Table 2.5. In the Figure, we added an horizontal line at 25% to help the inspection of the MAPEs. We conclude that for most periods, the MAPEs are under 25%, meaning that if we fitted the model for different estimation periods over the course of the study (January 2008 to August 2015) we could estimate the observed dengue case count for one or two weeks ahead with an error no more than 25%.

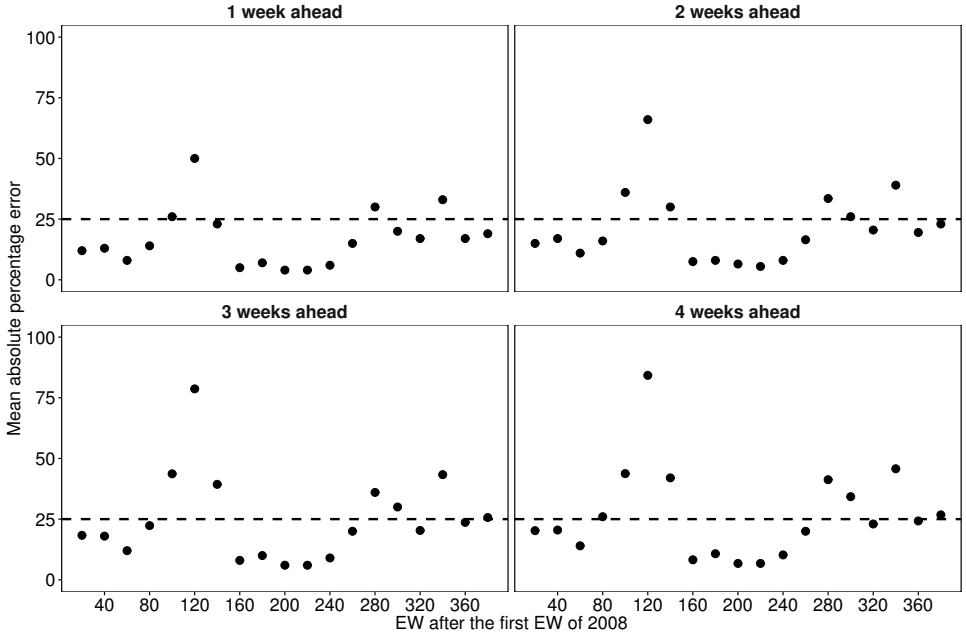


Figure 2.6: Posterior median of the MCMC simulations for the mean absolute percentage error (MAPE) to evaluate the short-term predictive performance of the final model in selected EWs after the first EW of January 2008

2.5 Discussion

In this report, DGLMs are employed to model time series of dengue disease case counts and meteorological variables. DGLMs for the data at hand included two components: the first substracts the temporal pattern, and the second models the covariate effect.
We observed weak time-varying associations between cases of dengue disease and solar radiation and temperature. Time-varying associations mean that the dengue case counts are associated with solar radiation and temperature changes over time, where some intervals show a positive association, while in other intervals the association is negative. DGLMs are a straightforward way to deal with count data, without the need to transform or alter the response variable, accounting for covariates with natural time-varying behavior.
For parameter estimation, we applied MCMC using WinBUGS 1.4, providing the flexibility to include constant and time-varying coefficients for calendar trend and covariates. There are few examples of studies including time-varying coefficients. Lee and Shaddick (2008) [44] fit DGLMs to pollution data and respiratory diseases, based on the block sampling algorithm from Knorr-Held (1999) [45]. Ruiz-Cardenas *et al.* (2012) [46] employed Integrated Laplace Approximation (INLA) to illustrate the fit of simulated and real time series of counts, using augmented data with the inclusion of time varying-coefficients for calendar trend and covariates.
Our findings can be summarized as follows: in the models without covariates, the best model was the RW1 TVCs (α) for trend. Within the models with CC (β_j) for covariates, we found the worst fit in models with CC (α) for trend, which display strong association (95% CIs not including zero) between weekly cases of dengue and temperature, solar radiation and rainfall, but not with relative humidity. However, models with RW1 or RW2 TVCs (α_t) for calendar trend had a good fit, revealing a weak association between dengue and the covariates. These findings are important because simple and multiple Poisson regression models with constant coefficients for the covariates are statistical methods commonly employed to model counts of infectious diseases like dengue [4].
For example, Hii *et al.* [16] modeled dengue and weather variables, applying a Poisson multiple regression model with piecewise linear spline functions for the covariates and constant coefficient terms to model auto regression, seasonality and trend. They validated the model by forecasting cases of dengue for week 1 of 2011 up to week 16 of 2012 using weather data alone.
In the class of models with RW1 TVCs ($b_{t,j}$) for the covariates, the best model corresponds to the simple dynamic Poisson model with RW1 TVCs (α_t) for calendar trend. After fitting the simple dynamic regression models, we fitted multiple dynamic regression models, with several combinations of TVCs ($b_{t,j}$) for the covariates, and we selected the model including all the meteorological variables. Our final model delineates the time-varying association between the covariates and cases of dengue, although the inspection of the mean estimates and 95% CIs of the RW1 TVCs ($b_{t,j}$) for the covariates shows a weak association.
In the literature associating dengue and weather variables, many of the modeling strategies

show strong association (evidenced by low p-values) between dengue and meteorological variables, with different lag periods. As an example, Xu *et al.* [19] established an association between absolute humidity (relative humidity adjusted by temperature) and dengue cases using a Poisson distributed lag non-linear model, with cubic splines for the covariates and accounting autoregression with constant coefficients for the lag-one and lag-two response.

We also evaluate the short-term predictive performance of the selected model, concluding that it enables relatively accurate ($< 25\%$ error) prediction of weekly dengue case counts at one or two weeks ahead although the predictions are strongly influenced by volatility in the weeks preceding the prediction periods, with high volatility associated with high MAPE in the predictions, as occurred in the peak of the 2010, 2013 and 2014 outbreaks in Bucaramanga.

Before finishing our discussion, we acknowledge some study limitations. The dengue case counts used in the data corresponded to the probable and confirmed cases reported to the official public health surveillance system in Colombia. The weekly dengue data was the sum of the the dengue and severe dengue cases per EW. Romero-Vega *et al.* (2014) [47] concluded that the expansion factor (the factor by which the reported cases should be multiplied to adjust for underreporting) of dengue was 7.6 for 2013, which is high. This implies that efforts to decrease underreporting must be undertaken to improve data quality for the entire surveillance system. It would be difficult to quantify the impact of underreporting in our conclusions, but still, the methods we used are valid for adjusted time series of dengue.

The covariates data (time series of temperature, rainfall, solar radiation and temperature) were a composition of several time series at daily and hourly temporal scales from several meteorological stations at different locations in the city. We summarized the data, averaging them for the different temporal scales and stations and consequently losing some data. However at some point the analyst must decide how to summarize the information to input variables for a modeling exercise. If the temporal scale is reduced (from weekly to daily data) the dengue case counts will be lower, and the Poisson models presented in the study could fit the data much better than Normal models.

One of this study's referees remarked on the absence of vector data in the study. We explored several sources of vector data in the city, but we did not find any data at the temporal scale of the study. We recognize that the inclusion of data for the distribution, presence and ecology of the vector would improve the conclusions of the study, but this is an opportunity to show that dengue in Colombia, and particularly in Bucaramanga, is a neglected disease, despite its huge impact on the population and the allocation of resources for dengue research (Villabona-Arenas *et al.*, 2016) [38].

One interesting experience in ongoing vectorial surveillance is in the city of Medellín, Colombia. Rúa-Uribe (2016) [48] reported that the Health Office of this city designed an entomological surveillance system using mosquito larval traps. We hope that the results of this interaction between the public sector and the research community will be disseminated to the country, and similar surveillance systems will be applied in all Colombian cities affected by arboviral diseases.

In the mean-time, for the city of Bucaramanga, we applied a dynamic Poisson model with

time-varying coefficients for the covariates and calendar trend, which helps to establish the association between climatic factors and dengue case counts at a small temporal scale, providing a prediction model within the bounds of the limitations presented in the study.
Forecasting models are commonly deployed in dengue research literature. Earnest *et al.* [10] compare the forecasting ability of the ARIMA model and the two-component Knorr-Held model (seasonal and epidemic Bayesian hierarchical time series model) to predict out-of sample cases of dengue. They found similar predictive ability (lower MAPE values) for the Bayesian K-H model and the ARIMA model.
Forecasting models of dengue disease usually account cyclical or seasonal behavior of the time series at hand. Earnest *et al.* [10] and Hii *et al.* [16] included seasonal trend by means of sinusoidal terms with trigonometric series structure. In a previous stage, we included seasonal terms, but we removed them from the models, allowing the time-varying coefficients for calendar trend alone account for dengue incidence trends.
We establish the short-term predictive performance of a model with time-varying coefficients (α_t) for calendar trend and time-varying coefficients ($b_{t,j}$) for meteorological covariates. We found a moderate predictive ability from the model to forecast cases of dengue disease at one or two weeks, which could be used by public health authorities interested in employing predictive models to help in the labors of dengue surveillance and control in Colombia.
For the future, we will explore the study models in different datasets from other cities of Colombia because, the environmental and physical conditions are generally similar between many cities and municipalities. The models presented in the study are not only available for use with climatic variables. They can also include data from vectorial studies, socioeconomic variables and many more, if these are available at weekly or monthly temporal scales. In conclusion, we found that dynamic generalized linear models can forecast dengue cases at one or two weeks in Bucaramanga, based on temperature, rainfall, solar radiation and relative humidity, and the models allow us to explore the association between weekly cases of dengue and these covariates through the time.

2.6 References

[1] Whitehorn J, Simmons CP. The pathogenesis of dengue. *Vaccine* 2011; **29**(42):7221–7228.

[2] Ocazionez RE, Cortés FM, Villar LA, Gómez SY. Temporal distribution of dengue virus serotypes in Colombian endemic area and dengue incidence. Re-introduction of dengue-3 associated to mild febrile illness and primary infection. *Memorias do Instituto Oswaldo Cruz* 2006; **101**(November):725–731.

[3] Naish S, Dale P, Mackenzie JS, McBride J, Mengersen K and Tong S Climate change and dengue: a critical and systematic review of quantitative modelling approaches *BMC Infectious Diseases* 2014; **14**:167 http://www.biomedcentral.com/1471-2334/14/167

[4] Imai C, Hashizume M. A Systematic Review of Methodology: Time Series Regression Analysis for Environmental Factors and Infectious Diseases. *Tropical Medicine and Health* 2015; **43**(1): 1–9.

[5] Racloz V, Ramsey R, Tong S, Hu W. Surveillance of Dengue Fever Virus: A Review of Epidemiological Models and Early Warning Systems. *PLoS Neglected Tropical Diseases* 2012; **6**(5):e1648.

[6] Runge-Ranzinger S, McCall PJ, Kroeger A, Horstick O. Dengue disease surveillance: an updated systematic literature review. *Tropical Medicine & International Health* 2014; **19**(9):1116–1160.

[7] Luz PM, Mendes BVM, Codeço CT, Struchiner CJ, Galvani AP. Time series analysis of dengue incidence in Rio de Janeiro, Brazil. *The American Journal of Tropical Medicine and Hygiene* 2008; **79**(6):933–939.

[8] Rúa-Uribe G, Suárez-Acosta C, Chauca J, Ventosilla P, Almanza R. Modelización del efecto de la variabilidad climática local sobre la transmisión de dengue en Medellín (Colombia) mediante análisis de series temporales[Modelling the effect of local climatic variability on dengue transmission in Medellin (Colombia) by means temporary series analysis]. *Biomédica* 2013; **33**(Supl. 1):9–11.[Spanish]

[9] Cho-Min-Naing A, Lertmaharit S, Khin-Saw-Naing A. Time-series analysis of dengue fever/dengue haemorrhagic fever in Myanmar since 1991. *Dengue Bulletin* 2002; **26**(1):24–32.

[10] Earnest A, Tan SB, Wilder-Smith A, Machin D. Comparing statistical models to predict dengue fever notifications. *Computational and Mathematical Methods in Medicine* 2012; **2012**:758.

[11] Martinez EZ, da Silva EAS. Previsão do número de casos de dengue em Ribeirão Preto , São Paulo , Brasil , por um modelo SARIMA. *Caderno de Saúde Pública* 2011; **27**(9):1809–1818.

[12] Martinez EZ, da Silva EAS, Fabbro ALD. A SARIMA forecasting model to predict the number of cases of dengue in Campinas, State of São Paulo, Brazil. *Revista da Sociedade Brasileira de Medicina Tropical* 2011; **44**(4):436–440.

[13] Gharbi M, Quenel P, Gustave J, Cassadou S, La Ruche G, Girdary L, Marrama L. Time series analysis of dengue incidence in Guadeloupe, French West Indies: forecasting models using climate variables as predictors. *BMC Infectious Diseases* 2011; **11**(1):166.

[14] Wongkoon S, Jaroensutasinee M, Jaroensutasinee K. Assessing the temporal modelling for prediction of dengue infection in northern and northeastern, Thailand. *Tropical Biomedicine* 2012; **29**(3):339–348.

[15] Wongkoon S, Jaroensutasinee M, Jaroensutasinee K. Development of temporal modeling for prediction of dengue infection in Northeastern Thailand. *Asian Pacific Journal of Tropical Medicine* 2012; **5**(3):249–252.

[16] Hii YL, Zhu H, Ng N, Ng LC, Rocklöv J. Forecast of Dengue Incidence Using Temperature and Rainfall. *PLoS Neglected Tropical Diseases* 2012; **6**(11):e1908.

[17] Lu L, Lin H, Tian L, Yang W, Sun J, Liu Q. Time series analysis of dengue fever and weather in Guangzhou, China. *BMC Public Health* 2009; **9**(1):395.

[18] Sang S, Gu S, Bi P, Yang W, Yang Z, Xu L, Yang J, Liu X, Jiang T, Wu H, *et al.*. Predicting Unprecedented Dengue Outbreak Using Imported Cases and Climatic Factors in Guangzhou, 2014. *PLOS Neglected Tropical Diseases* 2015; **9**(5):1–12.

[19] Xu HY, Fu X, Lee LKH, Ma S, Goh KT, Wong J, Habibullah MS, Lee GKK, Lim TK, Tambyah PA, *et al.*. Statistical modeling reveals the effect of absolute humidity on dengue in Singapore. *PLoS Neglected Tropical Diseases* 2014; **8**(5):e2805.

[20] Quintero-Herrera LLL, Ramírez-Jaramillo V, Bernal-Gutiérrez S, Cárdenas-Giraldo EVV, Guerrero-Matituy EAA, Molina-Delgado AHH, Montoya-Arias CPP, Rico-Gallego JAA, Herrera-Giraldo ACC, Botero-Franco S, *et al.*. Potential impact of climatic variability on the epidemiology of dengue in Risaralda, Colombia, 2010-2011. *Journal of Infection and Public Health* 2015; **8**(3):291–297.

[21] Huang X, Clements, Archie C, Williams G, Milinovich G, Hu W. A threshold analysis of dengue transmission in terms of weather variables and imported dengue cases in Australia. *Emerging Microbes and Infections* 2013; **2**(e87):1–7.

[22] Torres C, Barguil S, Melgarejo M, Olarte A. Fuzzy model identification of dengue epidemic in Colombia based on multiresolution analysis. *Artificial Intelligence in Medicine* 2014; **60**(1):41–51.

[23] Minh An DT, Rocklöv J. Epidemiology of dengue fever in Hanoi from 2002 to 2010 and its meteorological determinants. *Global Health Action* 2014; **7**(0):1–16.

[24] Ehelepola NDB, Ariyaratne K, Buddhadasa WMNP, Ratnayake S, Wickramasinghe M. A study of the correlation between dengue and weather in Kandy City , Sri Lanka (2003 -2012) and lessons learned. *Infectious Diseases of Poverty* 2015; **4**(42):1–14.

[25] Rodríguez J, Correa C. Predicción Temporal de la Epidemia de Dengue en Colombia: Dinámica Probabilista de la Epidemia.[Temporal prediction of the dengue epidemics in Colombia] *Revista de Salud Pública* 2009; **11**(3):443–453 [Spanish]

[26] Rodríguez Velásquez J, Vitery Erazo S, Puerta G, Muñoz D, Rojas I, Pinilla Bonilla L, Mora J, Salamanca D, Perdomo N. Dinámica probabilista temporal de la epidemia de dengue en Colombia. [Temporal probabilistic dynamics of the dengue epidemics in Colombia.] *Revista Cubana de Higiene y Epidemiología* 2011; **49**(1):74–83. [Spanish]

[27] Ferreira M, Gamerman D. Dynamic generalized linear models. *In: Dey, DK, Ghosh, SK, Mallick BK . Generalized linear models: a Bayesian perspective. Chapman & Hall/CRC Biostatistics Series*. 2000; 57–72.

[28] Schmidt AM, Pereira JBM. Modelling Time Series of Counts in Epidemiology. *International Statistical Review* 2011; **79**(1):48–69.

[29] Malhão TA, Casquilho Resende CM, Gamerman D, de Andrade Medronho R. Um modelo bayesiano para investigação de sobremortalidade durante epidemia de dengue na Região Metropolitana do Rio de Janeiro. *Caderno de Saúde Pública* 2013; **29**(March 2008):2057–2070.

[30] West M, Harrison PJ, Migon HS. Dynamic Generalized Linear Bayesian Models and Forecasting. *Journal of the American Statistical Association* 1985; **80**(389):73–83.

[31] West M, Harrison PJ. *Bayesian Forecasting and Dynamic Models*. 2 edn., Springer-Verlag: New York, 1997.

[32] Fahrmeir L, Tutz G. *Modelling Based on Generalized Linear Models*. Second edn., Springer-Verlag: New York, 2001.

[33] Chiogna M, Gaetan C. Dynamic generalized linear models with application to environmental epidemiology. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 2002; **51**(4):453–468.

[34] Shephard S, Pitt MK. Likelihood Analysis of Non-Gaussian Measurement Time Series. *Biometrika* 1997; **84**(3):653–667.

[35] Gamerman D. Markov chain Monte Carlo for dynamic generalised linear models. *Biometrika* 1998; **85**(1):215–227.

[36] Alves MB, Gamerman D, Ferreira MAR. Transfer functions in dynamic generalized linear models. *Statistical Modelling* 2010; **10**(1):3–40.

[37] Villar LA, Rojas DP, Besada-Lombana S, Sarti E. Epidemiological Trends of Dengue Disease in Colombia (2000-2011): A Systematic Review. *PLoS Neglected Tropical Diseases* 2015 **9**(3): e0003499. doi:10.1371/journal.pntd.0003499

[38] Villabona-Arenas CJ, Ocazionez Jimenez RE, Jimenez Silva CL. Dengue Vaccine: Considerations before Rollout in Colombia. *PLoS Neglected Tropical Diseases* 2016 **10**(6): e0004653. doi:10.1371/journal.pntd.0004653

[39] Padilla J, Rojas D, Sáenz-Gómez R Dengue en Colombia: epidemiología de la reemergencia a la hiperendemia [Dengue in Colombia: Epidemiology of Hyperendemic Reemergence] Guías de impresión (Bogotá, Colombia) 2012. ISBN: 9789584606617. [Spanish]

[40] Romero-Vega L. Vigilancia y Control en Salud Pública. Informe final del evento, Dengue, año 2012. [Surveillance and control in public health. Final report of the event, Dengue, year 2012] Instituto Nacional de Salud. Ministerio de Salud, Colombia[Spanish]

[41] Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine* 2009; **28**: 3049–3067.

[42] Gelman, A and Rubin, DB. Inference from iterative simulation using multiple sequences, *Statistical Science* 1992; **7**, 457-511.

[43] Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B(Statistical Methodology)* 2002; **64**(4):583–639.

[44] Lee D, Shaddick G. Modelling the effects of air pollution on health using Bayesian dynamic generalised linear models. *Environmetrics* 2008; **19**:785–804.

[45] Knorr-Held L. Conditional Prior Proposals in Dynamic Models. *Scandinavian Journal of Statistics* 1999; **26**(1):129–144.

[46] Ruiz-Cárdenas R, Krainski ET, Rue H. Direct fitting of dynamic models using integrated nested Laplace approximations - INLA. *Computational Statistics and Data Analysis* 2012; **56**(6):1808–1828.

[47] Romero-Vega L, Pacheco O, de la Hoz-Restrepo F, Díaz-Quijano FA. Evaluation of dengue fever reports during an epidemic, Colombia. *Revista de Saúde Pública* 2014; **48**, 899–905. DOI:10.1590/S0034-8910.2014048005321

[48] Rúa-Uribe, G Monitoreo entomológico por larvitrampas, una estrategia para apoyar la toma de decisiones en el control de ETV en Medellín [Entomological monitoring by larval traps, an strategy to support decision making in the control of vectorial transmitted diseases in Medellín]. Conference abstract. *Third International Dengue Integral Intervention Course. Bucaramanga, Colombia, August 10-13, 2016*. http://www.redaedes.org/english/index.php [Spanish]

2.7 Supplementary appendix

We employed the Gelman-Rubin (GR) diagnostic [1] to evaluate MCMC convergence by analyzing the difference between multiple Markov chains. Values substantially above 1 indicate lack of convergence. We show in Table 2.6 the GR statistics for the standard deviations (σ_α, σ_T, σ_{RF}, σ_{SR} and σ_{RH}) of models with RW1 time-varying coefficients α_t for calendar trend and RW1 time-varying coefficients for the covariates. For most

[1] Gelman, A and Rubin, DB. Inference from iterative simulation using multiple sequences, Statistical Science 1992; 7, 457-511.

models, the convergence of the selected parameters are close to the value of 1, while some of the GR statistics for σ_T and σ_{SR} display values slightly above 1. We accept those values, because the standard deviations for these covariates reflect the convergence of 395 parameters each, showing that few of the time-varying coefficients present difficulties to converge, but in general we accept them. Figure 2.7 shows the trace plots and densities of the standard deviations (σ_α, σ_T, σ_{RF}, σ_{SR} and σ_{RH}) from the model including all the covariates $b_{t,T} + b_{t,RF} + b_{t,SR} + b_{vRH}$. We observed more volatility in the trace-plots for σ_T and σ_{SR} than for σ_α, σ_{RF}, and σ_{RH}, with densities slightly skewed to the right.

Table 2.6: Gelman-Rubin diagnostic for the models with RW1 time-varying coefficients α_t for calendar trend and RW1 time-varying coefficients for the covariates

Model	σ_α	σ_T	σ_{RF}	σ_{SR}	σ_{RH}
$b_{t,T}$	1.00	1.19	-	-	-
$b_{t,RF}$	1.00	-	1.03	-	-
$b_{t,SR}$	1.00	-	-	1.02	-
$b_{t,RH}$	1.00	-	-	-	1.08
$b_{t,T} + b_{t,RF}$	1.00	1.03	1.03	-	-
$b_{t,T} + b_{t,SR}$	1.00	1.08	-	1.02	-
$b_{t,T} + b_{t,RH}$	1.00	1.32	-	-	1.03
$b_{t,RF} + b_{t,SR}$	1.01	-	1.02	1.14	-
$b_{t,RF} + b_{t,RH}$	1.00	-	1.01	-	1.00
$b_{t,SR} + b_{t,RH}$	1.00	-	-	1.03	1.01
$b_{t,T} + b_{t,RF} + b_{t,SR}$	1.00	1.03	1.01	1.18	-
$b_{t,T} + b_{t,RF} + b_{t,RH}$	1.00	1.04	1.00	-	1.04
$b_{t,T} + b_{t,SR} + b_{t,RH}$	1.01	1.10	-	1.02	1.02
$b_{t,RF} + b_{t,SR} + b_{t,RH}$	1.00	-	1.00	1.16	1.02
$b_{t,T} + b_{t,RF} + b_{t,SR} + b_{t,RH}$	1.01	1.04	1.00	1.29	1.01

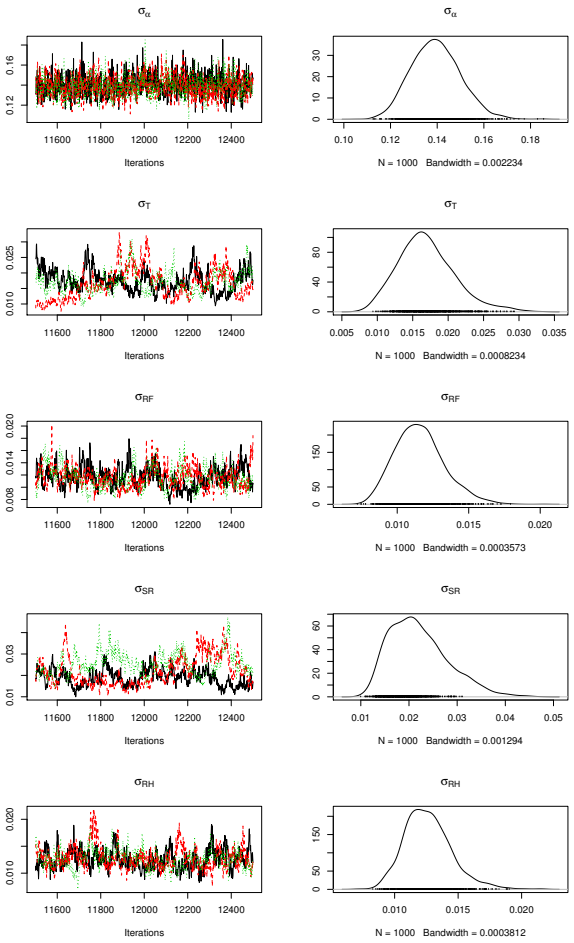


Figure 2.7: Trace plots and density plots for the standard deviations (σ_α, σ_T, σ_{RF}, σ_{SR} and σ_{RH}) of the selected model for inferences

Chapter 3

Relative risk estimation of dengue disease at small spatial scale

Abstract

Background: dengue is a high incidence arboviral disease in tropical countries around the world. Colombia is an endemic country due to the favorable environmental conditions for vector survival and spread. Dengue surveillance in Colombia is based in passive notification of cases, supporting monitoring, prediction, risk factor identification and intervention measures. Even though the surveillance network works adequately, disease mapping techniques currently developed and employed for many health problems are not widely applied. We select the Colombian city of Bucaramanga to apply Bayesian areal disease mapping models, testing the challenges and difficulties of the approach. Methods: we estimated the relative risk of dengue disease by census section (a geographical unit composed approximately by 1 to 20 city blocks) for the period January 2008 to December 2015. We included the covariates normalized difference vegetation index (NDVI) and land surface temperature (LST), obtained by satellite images. We fitted Bayesian areal models at the complete period and annual aggregation time scales for 2008-2015, with fixed and space-varying coefficients for the covariates, using Markov Chain Monte Carlo simulations. In addition, we used Cohen's Kappa agreement measures to compare the risk from year to year, and from every year to the complete period aggregation. Results: we found the NDVI providing more information than LST for estimating relative risk of dengue, although their effects were small. NDVI was directly associated to high relative risk of dengue. Risk maps of dengue were produced from the estimates obtained by the modeling process. The year to year risk agreement by census section was sligth to fair. Conclusion: the study provides an example of implementation of relative risk estimation using Bayesian models for

disease mapping at small spatial scale with covariates. We relate satellite data to dengue disease, using an areal data approach, which is not commonly found in the literature. The main difficulty of the study was to find quality data for generating expected values as input for the models. We remark the importance of creating population registry at small spatial scale, which is not only relevant for the risk estimation of dengue but also important to the surveillance of all notifiable diseases.
Keywords: Disease mapping; satellite images, Bayesian modeling, Cohen's Kappa

3.1 Introduction

Dengue is an arboviral disease characterized by fever and vascular complications and is endemic in many tropical and subtropical regions of the world [1]. Within the global efforts to control the disease, representation of the problem in space and time is key to supporting disease surveillance systems [2]. Spatial and spatio-temporal models are useful for generating risk maps for dengue disease, supporting early warning systems and intervention programs [2] [3].
Disease mapping tools correlated dengue incidence with socioeconomic, demographic and environmental variables using the Moran index in Brazil [4], while its geographical distribution has been characterized using spatial statistics and geographical information system (GIS) analysis in Ecuador [5]. Dengue disease mapping has also been combined with surveillance and monitoring of the *Aedes aegypti* vector in the Middle East [6], and in Peru, investigators have explored the association between dengue and clinical, meteorological, climatic, and sociopolitical variables through fuzzy association rule mining in a spatial setting [7]. At a micro-regional scale, Lowe *et al.* [8] included temperature, rainfall, the El Niño Southern Oscillation index, and other relevant socioeconomic and environmental variables in a spatio-temporal Bayesian hierarchical model implemented with Markov Chain Monte Carlo (MCMC), generating predictions at spatial and temporal levels and supporting a dengue alert system. Honorato *et al.* [9] studied the relationship between risk of dengue and sociodemographic variables using Bayesian spatial regression models in the municipalities of Espírito Santo, Brazil, while Ferreira and Smith [10] modeled the number of cases of dengue fever in Rio de Janeiro, Brazil, considering the cases as a Poisson random variable, with conditional autoregressive (CAR) priors in the spatial random effects, testing different neighborhood structures and covariates with fixed coefficients.
Colombia is highly endemic for dengue disease. From 2000 to 2011, the country experienced a stable annual incidence of dengue, with major outbreaks in 2001-2003 and 2010, followed by a considerable decrease of incidence in 2011, with cases mainly occurring in children ($<$ 15 years of age) and the highest incidence in 2009 in infants ($<$ 1 year of age) [11]. Small scale studies using dengue reports have investigated the spatial autocorrelation of dengue cases [12] and the association between dengue and satellite

environmental data using spatially stratified tests of ecological niche models [13]. Hagenlocher *et al.* [14] performed a spatial assessment of current socioeconomic vulnerabilities to dengue fever in 340 neighborhoods of a Colombian city through a spatial approach that included expert-based and purely statistical-based modeling of current vulnerability levels using a GIS.
At national level, Quintero *et al.* [15] used epidemiological surveillance data (weekly cases) and Poisson regression models to assess the influence of the El Niño Southern Oscillation index and pluviometry on dengue incidence, adjusting by year and week.At a regional scale, Cadavid-Restrepo *et al.* [16] explored the variation in spatial distribution of notified dengue cases in Colombia from 2007 to 2010, exploring associations between the disease and selected environmental risk factors through a Bayesian spatio-temporal conditional autoregressive model. The results elucidate the role of environmental risk factors in the spatial distribution of dengue disease, explaining how these factors can be used to develop and refine preventive approaches for dengue in Colombia. All these studies are strategic to the research of surveillance and control of dengue disease, demonstrating the importance of representation of the disease in space and time.
In Colombia, dengue epidemiological surveillance is based in passive notification of cases, coordinated by the 'Instituto Nacional de Salud' (Colombia National Institute of Health) [17], and supports monitoring, prediction, risk factor identification and intervention measures. Even though the surveillance network works adequately, and provides information to all the national institutions involved in dengue control, we appreciate that disease mapping techniques currently developed and employed for many health problems are not widely applied.
We selected the Colombian city of Bucaramanga to use Bayesian areal disease mapping models, testing the challenges and difficulties of the approach to dengue disease mapping. Bucaramanga was one of the cities with the highest dengue incidence in Colombia by year during the period 2008-2015. We estimated the relative risk of dengue disease, applying Bayesian spatial areal models to dengue case counts and satellite covariates (normalized difference vegetation index (NDVI) and land surface temperature (LST)) with fixed and space-varying coefficients, at a small spatial scale and with global and annual aggregation time scales, for the 2008-2015 period. Ideally, we would use data on the vector presence, distribution and ecology, but that kind of data does not exist for the city of Bucaramanga at the aggregation level of the study. We relied on satellite images to search associations between dengue incidence and environmental data. In addition, we provided a Bayesian model to estimate the Cohen's Kappa measure of agreement for the interpretation of the change in relative risk of dengue between the global and annual time scales.

3.2 Materials and Methods

3.2.1 Cases of dengue disease from Bucaramanga, Colombia

The city of Bucaramanga, Colombia is located at coordinates 7°7′07″ N - 73°06′58″ W, at 959 m above sea level. It covers an urban area of 27 km^2 and it has a population of 527,913 people in 2016, living in 220 neighborhoods nested in 17 communes. While Colombia presented an incidence rate of 436 cases per 100,000 persons in 2010, Bucaramanga reported an incidence rate of 1359.1 per 100,000 persons. We obtained data on incident cases of dengue disease (dengue and severe dengue) from the SIVIGILA (public health surveillance system) for the urban area of Bucaramanga for the period from January 2008 to December 2015.
We geocoded and allocated every case of dengue disease to one of 293 Bucaramanga census sections (geographical unit composed by approximately 20 closed city blocks) according to the cartography of the 2005 census from the national geostatistical framework of the 'Departamento Nacional de Estadística' (Colombia national statistics office) [18]. For geocoding purposes, we started with a database of 30,063 cases corresponding to the notified dengue cases from health institutions in Bucaramanga to the surveillance system. The cases were obtained from the database checked for duplicates reported to the surveillance system. The dengue cases data included address, sex, age and an identification code which anonymized the name and personal identification of the case to the geocoder. From this database, we selected the cases with address of residence belonging to Bucaramanga. We discarded cases without address, cases with rural address and wrong addresses. Then, an R [19] script sent batches of addresses to the web geocoding service of ArcGIS server. The returned geocoding were checked and accepted, or revised for a new geocoding cycle. At the end of the process, we succesfully geocoded to the urban area of Bucaramanga a total of 27,301 cases, which were aggregated to the spatial scale defined above, therefore our data does not relate to an identifiable natural person thus the data subject is not identifiable.
The cases aggregated by census section were aggregated along two time scales: a global scale, running for the entire study period (2008-2015); and an annual scale, resulting in eight respective datasets for each included year.
We obtained disaggregated data by census section, sex and five-years age groups from the 2005 census and calculated annual and global crude incidence rates according to these variables. We calculated expected values for dengue case counts by multiplying the global and annual crude incidence of dengue times the population by sex and age at census section level.

3.2.2 Satellite images for normalized difference vegetation index

We used satellite raster images obtained from Landsat Surface Reflectance (SR) Enhanced Thematic Mapper (ETM) 7, bands 3 and 4 (60 m resolution) for the years 2008, 2009, 2010, 2011; and from Landsat SR Operational Land Imager (OLI) and Thermal Infrared

Sensor (TIRS) 8, bands 4 and 5 (30 m resolution) for years 2013, 2014, and 2015 (Landsat Surface Reflectance products courtesy of the U.S. Geological Survey). We selected Landsat multispectral images based on those with the least cloud cover. Images covered the city of Bucaramanga and were taken from row 55, path 8 or path 7, with Universal Transversal Mercator (UTM) projection, zone 18 North and datum WGS-84. From Landsat SR ETM 7, we selected images from January 2, 2008; January 27, 2009; January, 14 2010; and February 2, 2011. From Landsat SR OLI-TIRS 8, we chose images from June 16, 2013; January 10, 2014; and January 4, 2015. We did not find any suitable Landsat images for the year 2012.

3.2.3 Satellite images for land surface temperature

We use Moderate Resolution Imaging Spectroradiometer (MODIS) satellite raster images from the MOD11A2 version 6 product [20] to obtain mean 8-day, per-pixel land surface temperature (LST), in a 1200 km x 1200 km grid. Each pixel value in the MOD11A2 is a simple average of all the corresponding MOD11A1 LST pixels collected within that 8-day period, with a pixel size of 1000 m x 1000 m. We selected the 'day time surface temperature' band.

3.2.4 Image processing

For the Landsat SR 7 ETM raster images, we calculated a composite NDVI raster image for the annual satellite images using band 4 (near infra red (NIR)) and band 3 (Red). For the Landsat SR 8 OLI-TIRS raster images, we calculated a composite NDVI raster image for the annual satellite images using band 5 NIR and band 4 (red). We applied the formula NDVI = (NIR band - red band)/(NIR band + red band), following Yuan *et al.* ([21]). Due to the absence of good quality images for 2012, we created a composite NDVI image for 2012 by pixel averaging of the composite NDVI images for 2011 and 2013.
To obtain an NDVI value by census section, we superimposed a mask comprised by the polygons (census sections) from the shape file of the city of Bucaramanga, onto the composite NDVI raster image, calculating the NDVI pixel mean by census section to the composite NDVI annual satellite images. We also produced a composite NDVI image at global aggregation scale (2008 - 2015 period) by pixel averaging by census section of all the composite NDVI images per year. For the composite NDVI image at global scale, we followed the same masking procedure to produce NDVI pixel mean by census section applied to the composite NDVI images by year.
For the MODIS LST raster images, we first reprojected the images from sinusoidal projection to UTM 18N projection, datum WGS84, and resampled to 30 m using the Modis reprojection tool (MRT) software [22]. We then created composite LST raster images per epidemiological period by pixel averaging all the reprojected and resampled images available in every epidemiological period. We generated composite LST images per year by pixel averaging all the composite LST raster images per epidemiological period. For every composite LST raster image per year, we applied a mask comprised

by the polygons (census sections) from the shape file of the city of Bucaramanga, and calculated the average LST by polygon (census section).
Finally, we created a composite LST image at global aggregation scale (2008-2015 period) by pixel averaging all the composite LST images per year. We produced LST average by census section in the composite LST image at global scale, following the same masking process to obtain LST average by census section per year. Raster image processing was done using the R software version 3.3 [19] with the `raster` package version 2.5-8 [23].

3.2.5 Statistical models

Disease mapping is the area of epidemiology that estimates the spatial pattern in disease risk over an extended geographical region in order to identify areas at high risk [24]. Besag *et al.* [25] developed a Bayesian hierarchical model for spatial analysis of areal data. Let θ_i be the log relative risk of a contagious disease with low transmission, O_i the observed number of cases and E_i the expected number of cases in area i. Assuming O_i are independent Poisson variables with mean $E_i\exp(\theta_i)$, where $\exp(\theta_i)$ is the relative risk of the disease and the linear predictor $\theta = t + u + v$ is adopted, where t is a term associated with measured covariates; and u and v are surrogates for unknown observed covariates. The u_i's represent variables with spatial structure and the v_i's represent spatially unstructured variables. In the hierarchical framework, prior probability distributions are assigned to u and v. For v, Normal prior with zero mean and variance λ, and for u,

$$p(u) \propto \left\{ -\sum_{i<j} \zeta_{ij}\phi(u_i - u_j) \right\}, u \in \mathscr{R} \tag{3.1}$$

where ϕ is a function of pairwise differences among u's, and ζ_{ij} are weights equal to zero for i non-contiguous to j. Besag *et al.* [25] consider two options for ϕ: $\phi(z) = z^2/2\kappa$ or $\phi(z) = |z|/\kappa$, where κ is an unknown constant. Choosing the first options for ϕ, the conditional structure of u follows

$$p(u|\kappa) \propto \frac{1}{\kappa^{n/2}} \left\{ -\frac{1}{2\kappa}\sum_{i\sim j} \zeta_{ij}(u_i - u_j)^2 \right\} \tag{3.2}$$

where $i \sim j$ represent neighbor areas, and the model is referred to as the 'Normal intrinsic autoregression.' For the non-zero ζ_{ij} several options are available, such as 1 for zones sharing a border and 0 otherwise, or the length of the boundary between contiguous zones [10]. The model with independent normal priors for v_i and Normal intrinsic priors for u_i is known as the 'convolution model'.
We fitted Poisson log normal models for relative risk, following Besag *et al.* [25], to the aggregated data at annual and global scale, with observation equation,

$$O_i \sim \text{Poisson}(E_i e^{\theta_i}) \tag{3.3}$$

where O_i, E_i and $\exp(\theta_i)$ are the observed count, the expected count, and the relative risk of dengue disease, respectively, in census section i ($i = 1, \cdots, m$, $m = 293$). For the linear

predictor, we explored the following structures:

$$\theta_i = \begin{cases} \alpha + u_i + v_i \\ \alpha + u_i + v_i + \beta_1 \text{NDVI}_i \\ \alpha + u_i + v_i + \beta_2 \text{LST}_i \\ \alpha + u_i + v_i + \beta_1 \text{NDVI}_i + \beta_2 \text{LST}_i \end{cases} \tag{3.4}$$

where the u_i are spatially correlated effects with Normal intrinsic conditional autoregressive (ICAR) priors distribution with precision parameter τ_u, and the v_i are the spatially uncorrelated effects with Normal prior distribution with zero mean and precision parameter τ_v. β_j ($j = 1,2$) are normally distributed fixed coefficients for NDVI and LST, with zero mean and precision 100. Uniform (0,1) hyperpriors were assigned to $\tau_u^{-1/2}$ and $\tau_v^{-1/2}$.

Next, we fitted models with spatially correlated effects with Leroux [26] Normal conditional autorregressive (CAR) prior distribution and fixed coefficients for the covariates,

$$\theta_i = \begin{cases} \alpha + w_i \\ \alpha + w_i + \beta_1 \text{NDVI}_i \\ \alpha + w_i + \beta_2 \text{LST}_i \\ \alpha + w_i + \beta_1 \text{NDVI}_i + \beta_2 \text{LST}_i \end{cases} \tag{3.5}$$

The w_i are the spatially correlated effects with Leroux Normal CAR prior distributions, with precision matrix τ_w with Gamma(1,0.001) hyperpriors. Finally, models were fitted with fixed and spatially varying coefficients for the covariates with Leroux CAR priors as follow:

$$\theta_i = \begin{cases} \alpha + (\beta_1 + b_{i,1})\text{NDVI}_i + (\beta_2 + b_{i,2})\text{LST}_i \\ \alpha + (\beta_1 + b_{i,1})\text{NDVI}_i \\ \alpha + (\beta_2 + b_{i,2})\text{LST}_i \end{cases} \tag{3.6}$$

where the $b_{i,j}$ are spatially varying coefficients for NDVI (j = 1) and LST (j = 2). For the linear predictor including two spatially varying coefficients $b_{i,j}$, these coefficients are modelled multivariate Normal with Leroux conditional mean vector $\mu_{i,j}$ ($j = 1,2$) and precision matrix $\xi_{2\times 2}$ following Congdon (2014) [27], where $\mu_{i,j}|\mu_{k\in\partial_i,j} = (\rho/(1 - \rho + \rho d_i))\sum_{k\in\partial_i}\mu_{k,j}$ and $\xi_{2\times 2} = (1-\rho+\rho d_i)\Xi_{2\times 2}$, where the precision matrix $\Xi_{2\times 2}$ is Wishart distributed with symmetric matrix S and 2 degrees of freedom.

The coefficient ρ establishes the degree of spatial structure of the spatial effects $\mu_{i,j}$. When ρ = 1, the Leroux prior for the spatial effects implies an Normal ICAR prior, while, ρ = 0, we have an independent model [28].

For the models with linear predictor including spatially varying coefficients for only one covariate, the $b_{i,1}$ or $b_{i,2}$ space-varying coefficients for NDVI or LST are modelled with Leroux conditional mean vector $\mu_{i,j}$ ($j = 1,2$) and precision τ_{b_j} (j = 1 for NDVI or j = 2 for LST), with Gamma(1,0.001) hyperpriors. All models included an intercept α with a diffuse improper prior .

From the Poisson log Normal models, we obtain the relative risk exp(θ_i) of dengue disease by census section and calculate point-wise mean estimates and 95% credible intervals

(CI). Choropleth maps were produced using the logarithm of the mean relative risk (θ_i) and the mean spatially correlated effects u_i by census section.
Additionally, the relative risk of dengue disease by census section was discretized as *low* or *high* risk, based on the lower bound of the 95% CI, where a value of 1 or less for the lower bound of the relative risk of dengue disease by census section indicates a *low* risk, and values exceding 1 signify a *high* risk. Choropleth maps of the discretized relative risk (DRR) of dengue disease are produced at global and annual aggregation scale.
Using the DRR of dengue disease, we calculated the Cohen's Kappa [29] coefficients for global-to-annual or annual-to-annual agreement of *low-high* risk by census section, using the following Bayesian model adapted from Lee and Wagenmakers [30]:

$$\begin{aligned}
\kappa &= (\delta - \psi)/(1 - \psi) \\
\delta &= \pi_1 + \pi_4 \\
\psi &= (\pi_1 + \pi_2)(\pi_1 + \pi_3) + (\pi_3 + \pi_4)(\pi_2 + \pi_4) \\
\mathbf{y} &\sim \text{Multinomial}([\pi_1, \pi_2, \pi_3, \pi_4], n) \\
\pi_i &\sim \text{Dirichlet}(1, 1, 1, 1)
\end{aligned}$$

where κ is the Kappa coefficient and $\mathbf{y}$ is the vector of counts in categories from the cross tabulation of *low* and *high* risk from global-to-annual or annual-to-annual agreement. Let year_a be one the eight study years and year_b other year not equal to year_a, the categories are as follows: y_1 = *low* risk in global scale or year_a and *low* risk in year_b; y_2 = *low* risk in global scale or year_a and *high* risk in year_b; y_3 = *high* risk in global scale or year_a and *low* risk in year_b; and y_4 = *high* risk in global scale or year_a and *low* risk in year_b.
We first calculated the global-to-annual agreement of DRR, between the pairs of DRR from model by global aggregation scale and the models for the data at annual aggregation scale. Second, we calculated the annual-to-annual agreement of DRR between all pairs of models fitted at annual aggregation scale. For the interpretation of the Kappa coefficients, we used the categories in Table 3.1, from Broemeling [31]. Models were fitted with

Table 3.1: Degree of agreement for Kappa

Degree of Agreement	**Poor**	**Slight**	**Fair**	**Moderate**	**Substantial**	**Perfect**
Kappa	<0	0-0.20	0.21-0.40	0.41-0.60	0.61-0.80	0.81-1.00

Markov Chain Monte Carlo (MCMC) using the WinBUGS software version 1.4 [32]. We utilize three chains with a burn-in period of 30,000 iterations, a final run of 10,000 iterations and thinning rate of 10, deriving a final sample of 1000 iterations by chain for the inference. To evaluate convergence, we check trace and density plots as well as the Gelman, Brooks and Rubin and Geweke tests [19]. Model selection was accomplished using the deviance, the deviance information criterion (DIC) and the number of effective parameters (p_D) [33].

3.3 Results

3.3.1 Summary statistics

Table 3.2 presents summary statistics for the dengue case counts, the standardized morbidity rate (SMR, the observed dengue cases divided by the expected dengue cases by census section), the NDVI, the LST and, the correlations between these variables, for the data aggregated at global and annual scales. For the global scale, a total of 27,301 cases

Table 3.2: Summary statistics for counts of dengue disease, NDVI and LST, by Bucaramanga census section, for globally and annually aggregated data, 2008-2015

Statistic	Global (2008-2015)	2008	2009	2010	2011	2012	2013	2014	2015
Dengue disease case counts									
Total	27301	1936	3131	6932	896	1546	4839	4956	3065
Min.	1	0	0	0	0	0	0	0	0
Max.	433	31	56	109	17	39	115	84	58
Mean	93.2	6.6	10.7	23.7	3.0	5.3	16.5	16.9	10.5
Standardized morbidity rate									
Min.	0.225	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Max.	4.052	8.409	6.346	5.401	12.294	5.979	4.683	5.129	5.502
Mean	1.098	1.125	1.082	1.143	1.193	1.089	1.096	1.101	1.165
NDVI									
Min.	0.135	0.091	0.130	0.096	0.134	0.144	0.142	0.146	0.129
Max.	0.792	0.767	0.813	0.757	0.803	0.850	0.896	0.784	0.779
Mean	0.368	0.346	0.394	0.355	0.368	0.379	0.390	0.367	0.346
LST									
Min.	28.8	29.3	28.8	27.6	28.4	29.1	29.3	28.9	28.5
Max.	34.0	34.2	34.9	32.9	34.2	34.5	34.6	34.8	32.8
Mean	32.1	32.3	32.5	30.7	31.8	32.4	32.2	32.9	31.7
Linear correlation									
Dengue-NDVI	-0.071	0.062	0.170	0.108	0.080	0.061	0.198	-0.014	0.086
Dengue-LST	0.127	0.016	-0.038	0.056	-0.063	-0.039	-0.126	0.017	-0.043
NDVI-LST	-0.629	-0.533	-0.640	-0.546	-0.583	-0.629	-0.568	-0.543	-0.522

(range by census section: 1 to 433) were reported and geocoded. The mean SMR for all census section was 1.098, with a minimum of 0.225 and a maximum of 4.05. The mean value of NDVI was 0.368, with a minimum of 0.135 and a maximum of 0.792. Mean LST for the aggregated data was 32.1°C, with a minimum of 28.8°C and a maximum of 34°C. Linear correlations between counts of cases of Dengue and NDVI (r = -0.071) and LST (r = 0.127) was weak, while, correlation between the NDVI and LST was moderate and negative (r = -0.629).

For the summary statistics at annual scale, 2010 was the year with the highest number of cases (n = 6932) followed by 2014 (n = 4956) and 2013 (n = 4839), with 2011 showing the lowest number of cases (896). The maximum number of cases in a census section ocurred in 2013 (n = 115), followed by 2010 (109) and 2014 (84). For the annual SMR, the year 2011 presented the highest average SMR (1.193) followed by 2010 (SMR = 1.143) and 2015 (SMR = 1.165).

The lowest maximum NDVI by census section corresponded to the year 2010 (NDVI = 0.757) followed, in order, by 2008 (NDVI = 0.767) and 2014 (NDVI = 0.784), while the

lowest mean NDVI were for years 2008 and 2015 (mean NDVI = 0.346) followed by 2010 (mean NDVI = 0.355).
With respect to the LST, the year 2010 displayed the lowest mean temperature (30.7°C), followed by 2015 (31.7°C) and 2011 (31.8°C), while for the rest of the years, the LST mean was close to 32°C.
The linear correlations between dengue case counts and NDVI or LST were low, whereas the most pronounced correlation was for 2013: between dengue and NDVI, r = 0.198; and between dengue and LST, r = -0.126. At an annual scale, the correlation between NDVI and LST was moderate and inverse for all years, with the strongest correlations for 2009 (r = -0.640) and 2012 (r = -0.629).

3.3.2 Model selection

Table 3.3 shows the selection statistics deviance, number of effective parameters (p_D) and DIC, for the models fitted at global and annual aggregation scales in Bucaramanga.

In general, the convolution models with CAR priors fitted the data better than the models with Leroux CAR priors, evidenced by the smallest deviance in the convolution models. For the data at global aggregation scale, the model with spatially correlated and uncorrelated effects with fixed coefficient for NDVI ($u_i + v_i + \beta_1$) presented the smallest DIC (DIC = 2364.6). For the year 2008, the smallest DIC was for the model with spatial effects with Leroux CAR priors and fixed coefficient for NDVI ($\beta_1 + w_i$) (DIC = 1457.8). For the years 2009, 2010, and 2014, we selected the models with correlated and uncorrelated spatial effects plus a fixed coefficient for NDVI ($u_i + v_i + \beta_1$), which presented the smallest DIC values of 1618.4, 1916.5, and 1798.3, respectively. For the years 2011, 2012, 2013, and 2015, the selected models contained space-varying coefficients for NDVI with Leroux CAR priors and a fixed coefficient for NDVI ($\beta_1 + b_{i,1}$), displaying the smallest DIC values of 1178.4, 1361.8, 1772.2, and 1629.1 respectively.

3.3.3 Parameter estimates of the selected model at global aggregation scale, 2008-2015

The selected model at global scale was the convolution model with fixed coefficient for NDVI ($u_i + v_i + \beta_1$). Figure 3.1 shows the map of the SMR logarithm, the map of the logarithm of the mean relative risk of dengue disease θ_i, the DRR of dengue disease, and the spatially correlated effects (u_i) from the model at global scale. The log mean relative risk of dengue disease shows high risk clusters in the south and north-west census sections of the city. The DRR map presents the areas where the 95% CIs for the relative risk do not include 1, and the map of spatial effects displays spatial correlation at the south of the city.

Table 3.3: Information criterion statistics, for relative risk models of dengue disease, 2008-2015

Model	Deviance	p_D	DIC	Deviance	p_D	DIC	Deviance	p_D	DIC
	Global scale 2008-2015				**2008**			**2009**	
$u_i + v_i$	2098.0	268.3	2366.3	1290.5	168.5	1459.0	1428.0	191.9	1619.9
$u_i + v_i + \beta_1$	**2097.2**	**267.4**	**2364.6**	1291.3	168.7	1460.0	**1427.7**	**190.7**	**1618.4**
$u_i + v_i + \beta_2$	2097.3	267.4	2364.7	1289.8	168.2	1458.0	1427.0	192.2	1619.2
$u_i + v_i + \beta_1 + \beta_2$	2097.8	267.9	2365.8	1289.5	168.7	1458.2	1427.9	191.7	1619.6
w_i	2098.1	268.9	2367.1	1296.2	163.3	1459.5	1428.8	196.0	1624.8
$w_i + \beta_1 + \beta_2$	2098.1	268.6	2366.7	1295.7	163.5	1459.1	1429.0	194.5	1623.5
$w_i + \beta_1$	2098.3	268.9	2367.2	**1294.5**	**163.3**	**1457.8**	1429.1	194.5	1623.6
$w_i + \beta_2$	2098.0	269.0	2367.0	1296.2	163.7	1459.9	1428.7	196.0	1624.6
$\beta_1 + b_{i,1} + \beta_2 + b_{i,2}$	2108.4	261.8	2370.2	1323.6	157.2	1480.8	1443.5	186.9	1630.5
$\beta_1 + b_{i,1}$	2113.8	257.0	2370.8	1346.2	140.6	1486.9	1465.5	175.7	1641.3
$\beta_2 + b_{i,2}$	2279.2	267.1	2546.3	1423.8	133.7	1557.5	1616.1	158.7	1774.8
		2010			**2011**			**2012**	
$u_i + v_i$	1684.3	233.5	1917.7	1063.3	120.5	1183.8	1198.4	164.8	1363.2
$u_i + v_i + \beta_1$	**1683.9**	**232.6**	**1916.5**	1063.5	120.6	1184.1	1199.6	165.6	1365.2
$u_i + v_i + \beta_2$	1684.4	233.8	1918.2	1063.0	119.9	1182.9	1198.2	165.6	1363.8
$u_i + v_i + \beta_1 + \beta_2$	1683.7	232.8	1916.5	1063.6	121.4	1185.0	1198.5	165.9	1364.5
w_i	1685.6	236.9	1922.5	1072.2	115.1	1187.3	1208.2	163.3	1371.4
$w_i + \beta_1 + \beta_2$	1686.0	236.0	1922.0	1071.0	117.4	1188.4	1207.7	163.6	1371.3
$w_i + \beta_1$	1686.0	236.5	1922.6	1072.2	115.9	1188.2	1208.4	163.3	1371.7
$w_i + \beta_2$	1684.0	236.2	1920.2	1071.8	115.9	1187.7	1207.8	163.8	1371.7
$\beta_1 + b_{i,1} + \beta_2 + b_{i,2}$	1694.3	227.6	1921.8	1065.9	121.0	1186.9	1206.8	156.6	1363.4
$\beta_1 + b_{i,1}$	1699.1	220.5	1919.6	**1077.6**	**100.7**	**1178.4**	**1214.9**	**146.8**	**1361.8**
$\beta_2 + b_{i,2}$	1823.0	228.0	2051.0	1168.5	69.3	1237.8	1332.8	119.6	1452.4
		2013			**2014**			**2015**	
$u_i + v_i$	1556.9	216.5	1773.4	1582.5	216.7	1799.2	1443.9	193.5	1637.4
$u_i + v_i + \beta_1$	1557.8	215.3	1773.1	**1581.9**	**216.4**	**1798.3**	1443.2	193.3	1636.5
$u_i + v_i + \beta_2$	1557.4	216.2	1773.6	1582.5	216.6	1799.2	1443.4	193.4	1636.8
$u_i + v_i + \beta_1 + \beta_2$	1557.7	215.4	1773.1	1581.8	217.3	1799.1	1443.4	193.8	1637.2
w_i	1560.0	216.5	1776.5	1586.4	220.0	1806.4	1447.1	197.5	1644.6
$w_i + \beta_1 + \beta_2$	1559.9	215.4	1775.3	1584.6	219.3	1803.9	1447.1	197.2	1644.3
$w_i + \beta_1$	1560.0	215.0	1775.0	1585.2	219.4	1804.7	1446.8	196.9	1643.7
$w_i + \beta_2$	1560.4	216.8	1777.2	1584.8	219.4	1804.2	1447.1	197.2	1644.3
$\beta_1 + b_{i,1} + \beta_2 + b_{i,2}$	1570.4	202.8	1773.1	1589.6	209.7	1799.3	1452.5	179.6	1632.1
$\beta_1 + b_{i,1}$	**1579.3**	**192.9**	**1772.2**	1597.2	202.3	1799.5	**1452.9**	**176.2**	**1629.1**
$\beta_2 + b_{i,2}$	1912.6	178.8	2091.4	1764.9	200.9	1965.8	1635.3	160.2	1795.5

3.3.4 Parameter estimates from models at annual aggregation scale, 2008-2015

Table 3.4 presents the mean point-wise parameters estimates and 95% CI from the selected models fitted at the annual aggregation scale. The selected model for 2008 was the model with spatially correlated effects with Leroux CAR priors plus fixed coefficient for NDVI ($w_i + \beta_1$). The point-wise mean estimate for the NDVI fixed coefficient is positive (0.294), and the 95% CI include zero (-0.214, 0.814), suggesting a weak association between dengue and NDVI by census section, at the same time, the point mean estimate of ρ is 0.486, which denotes moderate spatially correlated effects w_i in the relative risk of dengue disease.

The convolution model plus fixed coefficient for NDVI was the selected model for the years 2009, 2010, and 2014 ($u_i + v_i + \beta_1$). The point-wise mean and 95% CI estimates of the fixed coefficients for NDVI for years 2009, 2010, and 2014 show a strong, positive

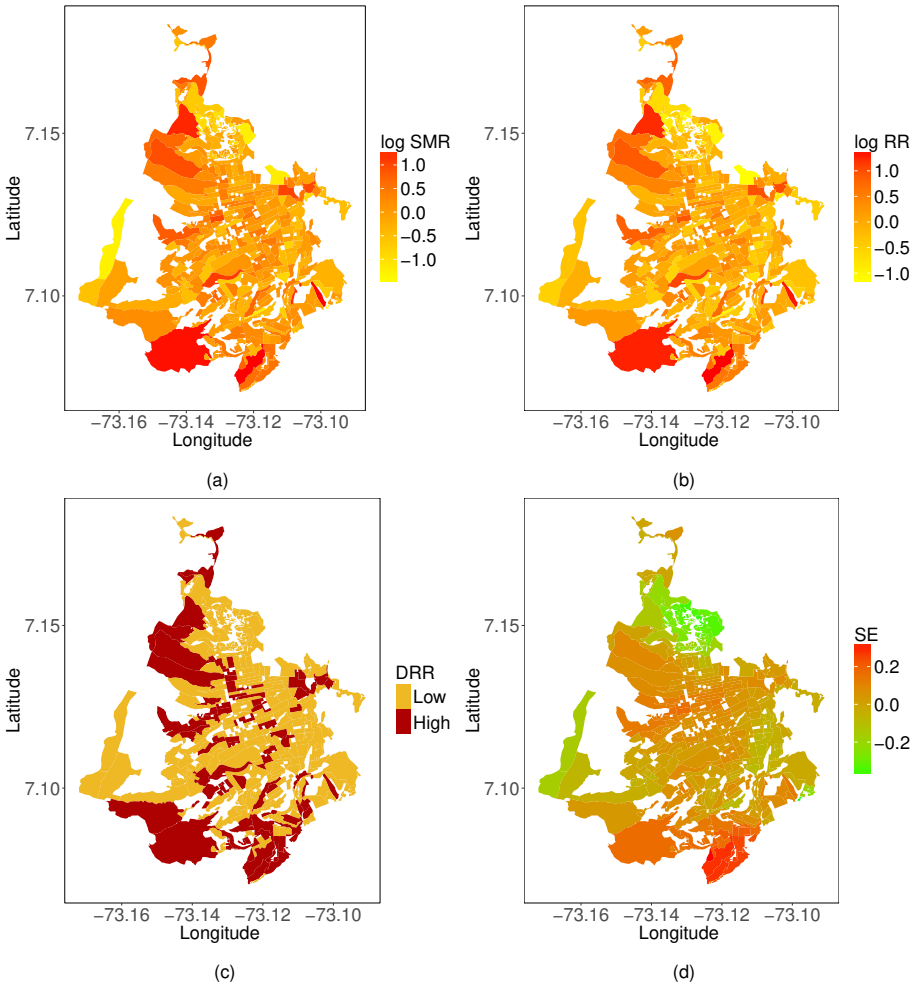


Figure 3.1: (a) Logarithm of the standardized morbidity rate (log SMR); (b) logarithm of the mean relative risk (log RR) of dengue disease; (c) discretized relative risk (DRR$_i$); and (d) mean spatial effects (SE) u_i, by census section for the data aggregated at global scale for the period 2008-2015

association between Dengue and NDVI for 2009 (0.589; 95% C.I: 0.168, 1.024), but not for 2010 and 2014. The mean of the spatially correlated effects w_i (2008) and u_i (2009, 2010 and 2014) are presented in Figure 3.2a. We observe the highest values of the mean spatial effects for the south of the city in 2008, but a scattered pattern of the mean of the spatially correlated effects for the rest of the years. The model with fixed coefficient and space-varying coefficients for NDVI was selected for years 2011, 2012, 2013, and 2014 ($\beta_1 + b_{i,1}$). The fixed coefficients for NDVI have to be accompanied by the space-varying coefficients, to be fully interpreted. The mean estimates for the fixed coefficient plus the space-varying coefficients for NDVI ($\beta_1 + b_{i,1}$) are presented in the Figure 3.2b. We discretized the space-varying coefficients in a similar way as the discretized relative risk

Table 3.4: Parameter estimates (point-wise mean and 95% CI) from the selected model at annual scale, 2008 - 2015 period

	Year			
Parameter	**2008**	**2009**	**2010**	**2011**
α	-0.173 (-0.399, 0.059)	-0.305 (-0.494, -0.115)	-0.19 (-0.363, -0.007)	0.108 (-0.12, 0.332)
σ_u	0.811 (0.658, 1.008)	0.323 (0.153, 0.512)	0.342 (0.199, 0.494)	
σ_v		0.457 (0.372,0.543)	0.45 (0.386,0.518)	
β_1	0.294 (-0.214, 0.814)	0.589 (0.168, 1.024)	0.43 (-0.041, 0.885)	-0.398 (-1.042, 0.249)
$\sigma_{b_{i,1}}$				1.399 (1.063, 1.835)
ρ	0.486 (0.217, 0.868)			0.062 (0.002, 0.236)
Parameter	**2012**	**2013**	**2014**	**2015**
α	0.116 (-0.11, 0.346)	-0.08 (-0.224, 0.071)	-0.093 (-0.28, 0.083)	0.049 (-0.138, 0.223)
σ_u			0.5 (0.346, 0.684)	
σ_v			0.404 (0.322, 0.483)	
β_1	-0.635 (-1.367, 0.12)	0.002 (-0.496, 0.502)	0.058 (-0.39, 0.536)	-0.222 (-0.867, 0.456)
$\sigma_{b_{i,1}}$	2.068 (1.601,2. 711)	1.785 (1.469,2. 188)		2.034 (1.666, 2.506)
ρ	0.279 (0.081, 0.592)	0.314 (0.146, 0.567)		0.316 (0.141, 0.575)

(DRR). We considered the association between the NDVI and dengue by census section to be *weak* when the lower bound of the 95% CI was 1 or less and, to be *strong* otherwise (Figure 3.2c). The discretized space-varying (DSV) NDVI coefficients enabled us to identify census sections where there was strong association between dengue incidence and NDVI. For 2012, 2013, and 2015, we observe a strong association in 4 to 8 census sections, while for 2011, the discretized effect was so low that there were no census sections showing an association between the covariate and dengue disease.

3.3.5 Mapping of relative risk of dengue disease, from models by annual scale

In this section, we begin by presenting the maps for the logarithm of the mean relative risk of dengue disease, and then we display the maps of the DRR of dengue, from the models selected by year 2008-2015.

Figure 3.3a presents the logarithm of the mean relative risk (Log RR) by year, for the period 2008 -2015. The smoothed estimates of the Log RR let us discern patterns of the disease distribution in Bucaramanga. We observe similar patterns for 2009, 2010, and 2011, with clusters of high relative risk in the southern and northwestern census sections. For 2008, 2012, 2013 and 2015, we observe slightly fewer zones with high relative risk. Estimates of relative risk of dengue for 2014 presented a great number of high-risk census sections in the northwestern, southern and central zones of the city. The maps of DRR of dengue disease help us to identify rapidly those census sections presenting the highest risk for each year (Figure 3.3b). The areas along the southern and northwestern edges of the city show a consistent tendency towards high relative risk, while the center generally presents a low risk, with some exceptions.

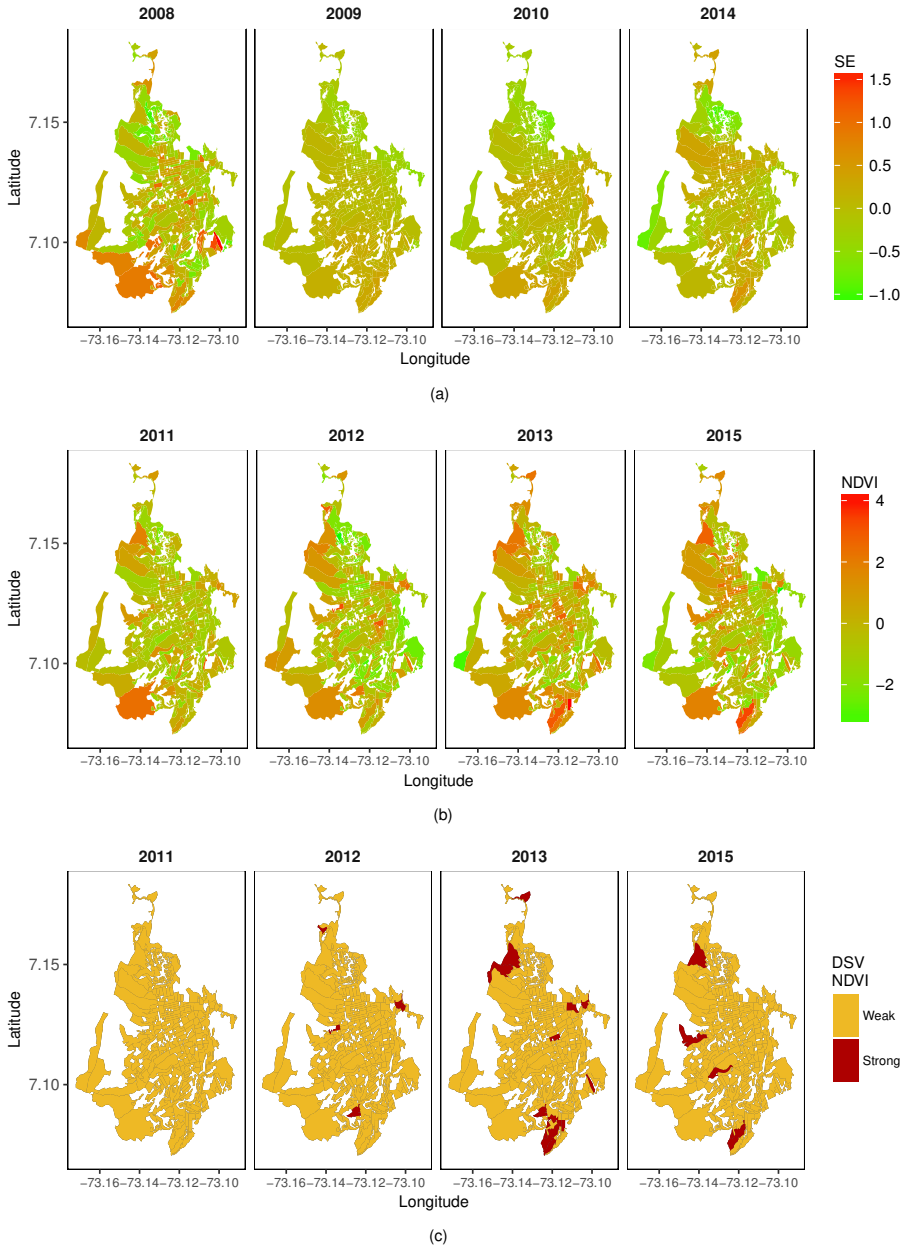


Figure 3.2: (a) Mean spatial effects (SE) w_i (2008) and u_i (2009, 2010, and 2014) from the selected models at annual aggregation scale; (b) mean space-varying NDVI coefficients ($\beta_1 + b_{i,1}$); and (c) discretized space-varying (DSV) NDVI coefficients ($\beta_1 + b_{i,1}$) for years 2011, 2012, 2013, 2015, from models at annual aggregation scale

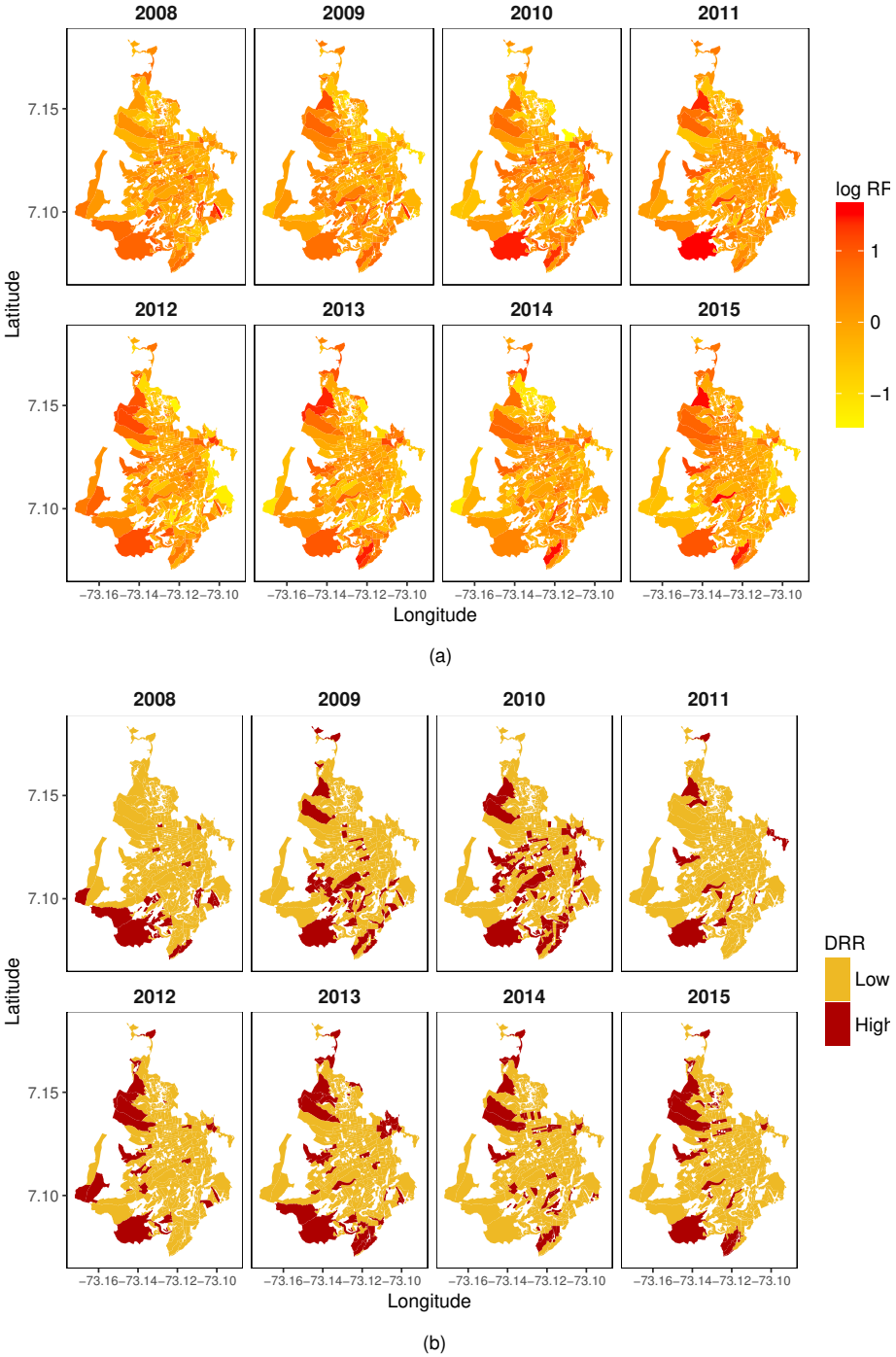


Figure 3.3: (a) Logarithm of the mean relative risk (log RR) of dengue disease , from models at annual aggregation scale 2008-2015; (b) discretized relative risk (DRR) of dengue disease, 2008-2015

3.3.6 Kappa coefficient for the global-to-annual and annual-to-annual agreement of DRR of dengue disease

We present estimates for the Kappa coefficient (point-wise mean and 95% CI) for global-to-annual and annual-to-annual agreement of DRR of dengue disease in Table 3.5. We interpreted the Kappa coefficient for agreement based on the coverage of the 95% CI over the categories of '*degree of agreement*' from Table 3.1. Firstly, we established the global-to-annual agreement of DRRs of dengue disease. Secondly, we determined the annual-to-annual agreement of the DRR of dengue disease.
From Table 3.5, we interpreted the global-to-annual agreement in DRR at years 2008, 2009, 2012, and 2015 to be *poor to fair*, while the agreement was *poor to slight* for 2011. Agreement of DRR of dengue disease was *slight to moderate* between the model at global scale and models for 2013 and 2014, and, *substantial to moderate* for the global scale model and the model for 2010.
For the annual-to-annual agreement of DRR, from 2009 to 2010, there was *poor to fair* agreement of DRR. From years 2011 to 2012 and 2012 to 2013, the agreement of DRR for dengue disease was *slight to moderate*. Additionally, the agreement was *slight to fair* from 2010 to 2011, 2013 to 2014, and 2014 to 2015. Finally, annual-to-annual agreement of DRR between other year pairs was almost always *poor to slight*.

Table 3.5: Kappa coefficient (point-wise mean and 95% CI) for global-to-annual and annual-to-annual agreement of discretized relative risk (DRR) of dengue by census sections, from models at global and annual aggregation scales in Bucaramanga, 2008-2015

		Annual scale						
Year	**Global scale**	**2009**	**2010**	**2011**	**2012**	**2013**	**2014**	**2015**
2008	0.231	0.146	0.107	0.087	0.142	0.115	0.003	0.042
	(0.124, 0.347)	(0.021, 0.288)	(0, 0.235)	(-0.027, 0.251)	(0.007, 0.303)	(-0.003, 0.255)	(-0.087, 0.126)	(-0.059, 0.173)
2009	0.327		0.271	0.207	0.207	0.226	0.141	0.097
	(0.207, 0.448)		(0.144, 0.402)	(0.078, 0.36)	(0.072, 0.355)	(0.094, 0.366)	(0.012, 0.281)	(-0.02, 0.234)
2010	0.513			0.143	0.158	0.294	0.134	0.075
	(0.397, 0.621)			(0.043, 0.261)	(0.046, 0.284)	(0.162, 0.425)	(0.013, 0.264)	(-0.035, 0.202)
2011	0.166				0.293	0.279	0.185	0.196
	(0.079, 0.271)				(0.116, 0.486)	(0.14, 0.431)	(0.054, 0.341)	(0.062, 0.357)
2012	0.254					0.363	0.225	0.269
	(0.144, 0.369)					(0.212, 0.513)	(0.085, 0.378)	(0.121, 0.424)
2013	0.466						0.219	0.231
	(0.352, 0.577)						(0.089, 0.361)	(0.1, 0.375)
2014	0.307							0.25
	(0.188, 0.428)							(0.112, 0.393)
2015	0.22							
	(0.099, 0.34)							

3.4 Discussion

In this study we applied Bayesian hierarchical models for spatial analysis of areal data to the estimation of relative risk of dengue disease in a Colombian city. We chose this particular city for its high incidence of dengue disease, in 2008-2015. We fitted models at global and annual scales for the study period. The hierarchical models included covariates (NDVI and LST) obtained from satellite raster images.

From the descriptive statistics, we found low correlation between the covariates and the dengue case counts by census section. NDVI and LST were moderately correlated by census sections at global and annual scales. Two main models were fitted: first the convolution model (spatially correlated effects with Normal ICAR priors, and uncorrelated spatial effects with Normal zero mean priors and precision τ_v); and second, spatially correlated effects model with Leroux Normal CAR priors. We used those structures with or without covariates. The covariates effects were modeled as fixed coefficients with Normal prior distributions with zero mean and high precision, or space-varying coefficients with Leroux Normal CAR priors.

We used MCMC to fit the models, and selected the models for inference, applying DIC measures. All the selected models (at global and annual scale) included NDVI fixed coefficients (2008, 2009, 2010, and 2015), and NDVI space-varying coefficients (2011, 2012, 2013, and 2014).

The convolution model was selected at global and annual scale for 2009, 2010, and 2014. The models with spatially correlated effects with Leroux Normal CAR priors were selected for 2008, 2011, 2012, 2013, and 2015. We illustrated the relative risk of dengue disease using two types of maps. First, we produced maps of the logarithm of the mean relative risk, containing smoothed estimates of relative risk, allowing us to distinguish clusters of high relative risk at global and annual scales. Second, we created maps of DRR of dengue disease, allowing us to identify zones where the risk is higher than in zones with 95% CIs including 1.

We employed satellite images from two sources (Landsat and MODIS) to calculate NDVI and LST data at areal level (census section). We were interested in establishing association between the information from raster images and dengue incidence at global and annual scale according to census section. The selected models for inferences did include NDVI but not LST. Parameter estimates (point-wise mean and 95% CI) for the association of the NDVI and dengue disease are mainly positive; however, they only reveal a strong association between dengue and NDVI, in 2009 (95% CI not including zero). Our results were different from some results in the literature. In Costa Rica, Troyo *et al.* [34] found negative coefficients for NDVI suggesting an inverse association with relative risk of dengue disease. They reported NDVI from satellite images at different resolutions, finding an association between high NDVI and low incidence of dengue. Meza-Ballesta *et al.* [35] reported the association between dengue incidence and high air temperature, high rainfall, and vegetation deterioration. Araujo *et al.* [36] analyzed dengue incidence in São Paulo, Brazil. They associated thermal remote sensing images, census data, and dengue incidence, and their findings point to low dengue incidence

in areas with high vegetation cover, and high incidence in areas with low vegetation coverage and land surface temperatures above 29 °C; however Nazri *et al.* [37] did not find NDVI to be a major factor influencing dengue incidence in Malaysia. Qi *et al.* [38] found that the associations between cases of Dengue disease and population density, GDP per capita, road density, and NDVI were nonlinear, and the risk of dengue disease declined gradually with rising NDVI.

These reports involving NDVI as a covariate did not use areal data, or the aggregation resolution was higher than the resolution used in our study. Additionally, at high spatial scale, the link between NDVI and rainfall variability in zones with epidemiological reports of dengue or malaria has been reported across South America [39], contributing to a better understanding of climate, environment, and epidemics facilitating the implementation of local and regional health early warning systems. In this context, our study contributes to understand the relationship between NDVI and incidence of dengue disease at a finer resolution.

As limitations of the satellite data employed in our study, we noted that for the period 2008-2011, we had to impute the NDVI in some census sections because of known issues with the sensor from Landsat ETM 7. We employed a convolution model to make the imputation of the missing NDVI values. For the NDVI data, the pixel resolution was 30 m, while the pixel resolution of the MODIS LST data was 1000 m, resampled to 30 meters, which makes some census section highly associated with respect to the LST. Additionally, for LST, we used single raster image per year, which is a composite of around 48 MODIS images. Moreover, finding high-quality Landsat images for the period was difficult, mainly because of cloud cover. In addition, we used a non-conventional method to treat information from satellite images, as compared to the literature, because we employed continuous variables such as NDVI and LST in space obtaining areal inputs as covariates, using the mean value for every census section.

With respect to the data quality on dengue, we employed official data, aggregating the cases into census sections acknowledging factors such as age or sex by means of internal standardization. We consider our dengue case counts to be high-quality data given that the surveillance system is based on compulsory reporting physicians in Colombia at all service levels, but keeping in mind critic considerations regarding possible underreporting as shown by Romero-Vega *et al.* [40].

For the spatial aggregation level and census data, we based our population estimates on the 2005 census, with the assumption that city's population has been fairly stable throughout the study period. We recognize the possible bias produced by the use of 2005 data for estimation of the expected values for the study period. This is one of the challenges of the study, but it is not only for the disease under study. At the time of the realization of this study, there were not updated data for population living in Bucaramanga, at census section level. The Colombian government updates its census every ten years. It was programmed a census for 2016, but was not developed. We consider this finding extremely important, if public health authorities are interested to provide information, not only for temporal analysis (as is currently done) but also for spatial analysis at finer resolution scales (census block, section or sector). This study provides the space to recommend to the Colombian authorities, to build real-time registries of population, which support evaluation, decision

making and intervention, not only for dengue disease, but for all the notifiable diseases. Additionally, we are aware that the spatial aggregation scale changes the conclusion from areal data as shown by Khormi and Kumar [41], as does the choice of neighborhood structure [10]. The spatial aggregation scale used in this study corresponds to the spatial scale supplied by the official cartography from the 2005 census.
Together with the generation of relative risk maps and the evaluation of satellite data associated with incidence of dengue disease, we determined whether there was an association in the DRR of dengue disease by census section from year to year, and from the risk at global scale and annual scale. To this end, we employed the Kappa coefficients to define the global-to-annual agreement of risk when the relative risk is discretized as *low* or *high* risk, estimating the Kappa coefficients, using a Bayesian model. We have found substantial to moderate agreement between the DRR at global scale and the year 2010, and moderate to fair agreement for 2013 and 2014. For the rest of the global-to-annual agreement or all the annual-to-annual agreement of the DRR for consecutive years, we found slight to fair agreement, reflecting a volatile pattern of dengue risk by census section from year to year. The main results for the annual-to-annual agreement of the discretized relative risk for non consecutive year, point to low agreement between census sections in terms of risk level between years.

3.5 Conclusions

We applied disease mapping models at small spatial scale in a Colombian city with fixed and space-varying coefficients for covariates derived of satellite images. We found the NDVI associated to high dengue risk by census section. The modeling process produced relative risk maps of dengue disease, allowing us to identify areas with high risk in the city. We compared the concordance of high risk by census section between the global aggregated model and the annual aggregated models, and between years in the 2008-2015 period. We found in general, slight to fair agreement in high risk. The information obtained by the use of disease mapping models is not currently available to public health authorities in Colombia. We highlight the importance of transforming raw spatial data into relative risk maps of dengue disease for planning and implementation of public health strategies. The map quality depends on high-quality population data at the selected spatial aggregation scale, which are sometimes difficult to obtain; and the geocoding process to allocate every case to a correct spatial coordinate. Information from satellite images improves the output of the spatial modeling, by associating environmental variables to the dengue incidence. For future research we are interested to apply disease mapping models in similar datasets from municipalities in Colombia. The main limitation is the cost of geocoding the data from the official records in terms of work-time and accurate geocoding. However, after the geocoding task is finished, the data will be available not only to spatial analysis, but also to temporal and spatiotemporal analysis, including the covariates with constant or time-varying coefficients [42]. We would like to present the method to the public health authorities from diverse surveillance offices in Colombia, and

test the field applicability of the disease mapping models discussed in our study. Finally, we also are interested to apply disease mapping methods based on hierarchical Bayesian models to map the relative risk not only for dengue disease, but also for arboviral diseases like Chykungunya and Zika virus disease at small spatial scale in Colombia, to support public health authorities in disease surveillance and control strategies.

3.6 References

[1] Whitehorn, J.; Simmons, C.P. The pathogenesis of dengue. *Vaccine*. 2011; **29**(42), 7221–7228

[2] Louis, V.R., Phalkey, R., Horstick, O., Ratanawong, P., Wilder-Smith, A., Tozan, Y., Dambach, P. Modeling tools for dengue risk mapping - a systematic review. *International Journal of Health Geographics*. 2014; **13**(1): 50

[3] Racloz, V., Ramsey, R., Tong, S., Hu, W. Surveillance of dengue fever virus: A review of epidemiological models and early warning systems. *PLoS Neglected Tropical Diseases*. 2012; **6**(5): 1648

[4] Mondini, A., Chiaravalloti-Neto, F. Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a Brazilian city. *Science of The Total Environment*. 2008; **393**(2-3): 241–248

[5] Castillo, K.C., Körbl, B., Stewart, A., Gonzalez, J.F., Ponce, F. Application of spatial analysis to the examination of dengue fever in Guayaquil, Ecuador. *Procedia Environmental Sciences*. 2011; **7**: 188–193

[6] Khormi, H.M., Kumar, L., Elzahrany, R.A. Modeling spatio-temporal risk changes in the incidence of dengue fever in Saudi Arabia: A geographical information system case study. *Geospatial Health*. 2011; **6**(1): 77–84

[7] Buczak, A.L., Koshute, P.T., Babin, S.M., Feighner, B.H., Lewis, S.H. A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC Medical Informatics and Decision Making*. 2012; **12**: 124

[8] Lowe, R., Bailey, T.C., Stephenson, D.B., Graham, R.J., Coelho, C.A.S., S Carvalho, M., Barcellos, C. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers and Geosciences*. 2011 **37**(3): 371–381

[9] Honorato, T., Lapa, P.P.D.A., Sales, C.M.M., Reis-Santos, B., Tristão-Sá, R., Bertolde, A.I., Maciel, E.L.N. Spatial analysis of distribution of dengue cases in Espírito Santo, Brazil, in 2010: use of Bayesian model. *Revista Brasileira de Epidemiologia*. 2014; **17**(Suppl D.S.S.): 150–159

[10] Ferreira, G., Schmidt, A. Spatial modelling of the relative risk of dengue fever in Rio de Janeiro for the epidemic period between 2001 and 2002. *Brazilian Journal of Probability and Statistics*. 2006; **20**: 29–47

[11] Villar, L.A., Rojas, D.P., Besada-Lombana, S., Sarti, E. Epidemiological trends of dengue disease in Colombia (2000-2011): a systematic review. *PLoS Neglected Tropical Diseases*. 2015; **9**(3): 1–16.

[12] Londoño C, L.A., Restrepo, C.E., Marulanda, E.O., Distribución, M.E. Spatial distribution of dengue based on Geographic Information Systems tools, Aburra Valley. *Revista de la Facultad Nacional de Salud Pública*. 2014; **32**(1): 7–15.

[13] Arboleda, S., Jaramillo-O., N., Peterson, A.T. Mapping environmental dimensions of dengue fever transmission risk in the Aburrá Valley, Colombia. *International Journal of Environmental Research and Public Health*. 2009; **6**(12): 3040–3055.

[14] Hagenlocher M, Delmelle E, Casas I, Kienberger S. Assessing socioeconomic vulnerability to dengue fever in Cali, Colombia: statistical vs expert-based modeling. *International Journal of Health Geographics*. 2013; **12**(1): 36.

[15] Quintero-Herrera LL, Ramírez-Jaramillo V, Bernal-Gutiérrez S, Cárdenas-Giraldo EV, Guerrero-Matituy EA, Molina-Delgado AH, Montoya-Arias CP, Rico-Gallego JA, Herrera-Giraldo AC, Botero-Franco S., Rodríguez-Morales A.J.: Potential impact of climatic variability on the epidemiology of dengue in risaralda, colombia, 2010-2011. *Journal of Infection and Public Health*. 2015; **8**(3): 291–297.

[16] Cadavid-Restrepo A, Baker P, Clements ACA.: National spatial and temporal patterns of notified dengue cases, Colombia 2007-2010. *Tropical medicine & international health*. 2014; **19**(7): 863–871.

[17] Zambrano P.: Protocolo de Vigilancia en Salud Pública, Dengue (Surveillance Protocol in Public Health, Dengue). Instituto Nacional de Salud (National Institute of Health), Santafé de Bogota, Colombia (2014). Instituto Nacional de Salud (National Institute of Health). http://www.ins.gov.co/lineas-de-accion/Subdireccion-Vigilancia/sivigila/Paginas/protocolos.aspx

[18] Departamento Administrativo Nacional de Estadística, g.o. Dirección de Geoestadística (National Administrative Department of Statistics: Capa del Nivel de Sector Urbano (urban Sector Level Layer). (2005). Marco Geoestadístico Nacional (National Geostatistical Framework). http://www.dane.gov.co/

[19] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. 2016. R Foundation for Statistical Computing. https://www.R-project.org/

[20] Wan Z, Hook S, Hulley G. MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006. NASA EOSDIS Land Processes DAAC, 2015, (2015). NASA EOSDIS Land Processes DAAC, 2015. https://doi.org/10.5067/MODIS/MOD11A2.006

[21] Yuan F, Bauer ME.: Comparison of impervious surface area and normalized difference vegetation index as indicators of surface urban heat island effects in Landsat imagery. *Remote Sensing of Environment*. 2007; **106**(3): 375–386

[22] United States Geological Service: Modis Reprojection Tool User's Manual. Release 4.1 April 2011. Land Processes DAAC. USGS Earth Resources Observation and

Science, (2011). Land Processes DAAC. USGS Earth Resources Observation and Science

[23] Hijmans RJ, van Etten J.: Raster: Geographic Analysis and Modeling with Raster Data. (2016). R package version 2.5-8. `http://CRAN.R-project.org/package=raster`

[24] Lee D.: A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology*. 2011; **2**(2); 79–89

[25] Besag J, York J, Mollie A.: Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*. 1991; **43**(1): 1–59

[26] Leroux BG, Lei X, Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: M, H., Berry.D. (eds.) Statistical Models in Epidemiology, the Environment and Clinical Trials, Springer, New York; 1999. pp. 179–191.

[27] Congdon P.: Applied Bayesian Modelling, 2nd edn., p. 464. John Wiley & Sons Ltd, West Sussex, England (2014)

[28] Banerjee S, Carlin B, Gelfand A. Hierarchical Modeling and Analyisis for Spatial Data. Chapman & Hall/CRC biostatistics series, Boca Raton, FL. 2015.

[29] Cohen J.: A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*. 1960; **20**: 37–46

[30] Lee MD, Wagenmakers EJ.: Bayesian Cognitive Modeling, A Practical Course. Cambridge University Press: University Printing House, Cambridge CB2 8BS, United Kingdom. 2014.

[31] Broemeling LD.: Bayesian Methods for Measures of Agreement. Chapman & Hall/CRC biostatistics series, Boca Raton, FL. 2009.

[32] Lunn D, Spiegelhalter D, Thomas A, Best N.: The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine*. 2009; **28**: 3049–3067

[33] Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A.: Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B(Statistical Methodology)*. 2002; **64**(4): 583–639

[34] Troyo A, Fuller DO, Calderón-Arguedas O, Solano ME, Beier JC.: Urban structure and dengue fever in Puntarenas, Costa Rica. *Singapore Journal of Tropical Geography*. 2009; **30**(2): 265–282

[35] Meza-Ballesta A, Gónima L. Influencia Del Clima Y De La Cobertura Vegetal En La Ocurrencia Del Dengue (2001-2010). *Revista Salud Pública*. 2004; **16**(2): 293–306

[36] Araujo RV, Albertini MR, Costa-da-Silva AL, Suesdek L, Franceschi NCS, Bastos NM, Katz G, Cardoso VA, Castro BC, Capurro ML, Allegro VLAC.: São Paulo urban heat islands have a higher incidence of dengue than other urban areas. *Brazilian Journal of Infectious Diseases*. 2015; **19**(2): 146–155

[37] Nazri C, Hashim A, Rodziah I: Distribution pattern of a dengue fever outbreak using GIS. *Journal of Environmental Health Research*. 2009; **9**(2002): 1–10

[38] Qi X, Wang Y, Li Y, Meng Y, Chen Q, Ma J, Gao G.: The effects of socioeconomic and environmental factors on the incidence of dengue fever in the pearl river delta, china, 2013. *PLoS Neglected Tropical Diseases*. 2015; **10**: 0004159

[39] Tourre YM, Jarlan L, Lacaux JP, Rotela CH, Lafaye M.: Spatio-temporal variability of NDVI–precipitation over southernmost South America: possible linkages between climate signals and epidemics. *Environmental Research Letters*. 2008; **3**: 044008

[40] Romero-Vega L, Pacheco O, de la Hoz-Restrepo F, Díaz-Quijano FA.: Evaluation of dengue fever reports during an epidemic, Colombia. *Revista de Saúde Pública*. 2014; **48**(6): 899–905

[41] Khormi HM, Kumar L.: The importance of appropriate temporal and spatial scales for dengue fever control and management. *Science of the Total Environment*. 2012; **430**: 144–149

[42] Martínez-Bello D, López-Quílez A, Torres-Prieto A.: Bayesian dynamic modeling of time series of dengue disease case counts. *PLoS Neglected Tropical Diseases*. 2017; **11**(7): 0005696

Chapter 4

Spatiotemporal modeling of relative risk of dengue disease in Colombia

Abstract

Spatiotemporal modeling of relative risk of dengue disease provides useful risk maps for surveillance and forecasting. The objective of the study was to generate smoothed estimates of relative risk applying hierarchical Bayesian spatiotemporal models, including covariates derived from satellite images containing land surface temperature (LST) and normalized difference vegetation index (NDVI), for the period January 2009–December 2015, in a medium–sized Colombian city. Our models are based on the spatiotemporal interaction modeling framework of relative risk, where the interaction effects are unstructured, temporal, spatial or inseparable, at a small spatial and temporal scales. We fitted the models using Markov chain MonteCarlo Simulations (MCMC), selecting the best model using leave-one-out (LOO) cross-validation and widely applicable information criteria (WAIC). Our best model was the inseparable spatiotemporal interaction-effects plus LST with constant coefficient model. We found a weak, positive association between LST and cases of dengue. We discussed the strengths and weaknesses of our spatiotemporal models given the spatial and temporal resolution selected in the study.
Keywords: space-time interaction effects model, areal data, normalized difference vegetation index, land surface temperature.

4.1 Introduction

Dengue is a mosquito-borne pandemic-prone, viral disease, affecting poor urban areas, city outskirts and crowded neighborhoods in tropical and subtropical countries. Every year, there are an estimated 50-100 million new infections in over 100 countries where dengue is endemic. The virus has four distinct serotypes (DEN–1, DEN–2, DEN–3 and DEN–4) belonging to the genus *Flavivirus*, family *Flaviviridae*. Dengue disease is the leading cause of hospitalization and death in children and adults in low- and middle-income countries [1].

Louis *et al.* ([2]) reviewed the spatial and temporal epidemiological factors associated with the risk of dengue disease, concluding that surveillance and prediction risk maps are powerful tools to improve decision making in public health; however, these tools require diverse, good quality data with adequate spatial and temporal resolution. Methods of dengue risk mapping are based on data aggregated at several spatial or spatiotemporal scales, following statistical methods like the Knox test [3] to find clusters of dengue cases using point-data, the global and local spatial autocorrelation index Moran's I statistic [4], the Getis G_i(d) statistic [5] and local indicators of spatial association LISA [6]. Clusters and hotspot detection that cannot be explained by the assumption of spatial or space-time randomness can be subjected to analysis with the space-time scan statistics [7]. Spatiotemporal dengue data and environmental and socioeconomics factors have been associated applying kernel density estimation and inverse distance weighting [8], k-means clustering algorithm [9], geographically weighted regression models, stochastic Bayesian Maximum entropy [10], semiparametric Bayesian Poisson space-time structured additive models [11], Poisson and negative binomial relative risk models for disease mapping and mapping reconstruction [12], among other techniques. The Knox test has been applied to dengue data from Argentina and Brazil [13] and Morato [14]; and the widely applied space-time scan statistics to dengue data from Vietnam [15]), Saudi Arabia [16], Malaysia [17], Sri Lanka [18], Nepal [19], Taiwan [20] and Argentina [21]. Global and local autocorrelation indexes have been applied in Brazil [22], Taiwan [23] and Ecuador [24]. Kernel density estimation and inverse distance weighting have been applied to dengue data from Brazil [25], [26] Taiwan [27], India [28] and Mexico [29]; k-means clustering have been employed to data from Taiwan [30]; and stochastic Bayesian maximum entropy applied to dengue data from Taiwan [31–33].

Relative risk estimation models represent the excess risk in small areas compared to a given background risk rate [34]. Poisson relative risk models have been applied to dengue data from Brazil [35], Australia [36], Colombia [37] and Indonesia [38], while negative binomial relative risk models have been fitted to dengue data from Brazil [39], [40], Ecuador [41] and Thailand [42].

Knorr-Held [43] proposes interaction-effects (IE) models to represent spatiotemporal variation of disease risk, extensible to relative risk estimation. Knorr-Held presents four types of interaction effects models: unstructured, temporal, spatial and inseparable, ranging from complete independence to full dependence. A careful look into the dengue literature reveals that the only model exploring spatiotemporal interactions-effects of

relative risk includes unstructured interaction effects [38].
Colombia is among the countries with the highest incidence of dengue disease [44]. The period from 2000 to 2011 was characterized by a stable 'baseline' annual number of dengue cases, with major outbreaks in 2001-2003 and 2010, followed by a decrease in 2011. Disease control has been difficult, prompting Colombia to join other countries testing dengue vaccines [45], despite criticism from a part of the research community regarding the lack of epidemiological and ecological information about the disease [46]. In Colombia, relative risk disease mapping for relative risk of dengue disease is not currently used as an analytical tool supporting public health surveillance and control programs for dengue. Cadavid *et al.* [37] applied geographical information systems (GIS) and Bayesian spatial Poisson relative risk models to describe the spatiotemporal patterns of notified dengue incidence in Colombia between 2007 and 2010. They identified dengue clusters at the municipality level, generating smoothed patterns of dengue risk according to its association with environmental temperature, precipitation and elevation to the distribution and dynamics of the disease.
The aim of this study is to explore Bayesian Poisson spatiotemporal IE models of relative risk of dengue disease, using data from a medium-sized Colombian city severely affected by dengue disease outbreaks, from January 2009 to December 2015. In addition, the association between dengue cases and environmental factors obtained from satellite images is explored using the Bayesian spatiotemporal models.
The paper is organized as follows. Section 2 presents the dengue data, the satellite covariates and the image processing to obtain the covariates. Section 3 discusses the spatiotemporal modeling framework of risk estimation. Section 4 discusses the spatiotemporal models applied to the dengue data and covariates. In Section 5, we present the results of the inference, and in Section 6, we discuss our findings.

4.2 Data

4.2.1 Dengue disease case counts

The city of Bucaramanga is located in the north easthern part of Santander Province, Colombia (7°07'07"N, 73°06'58"W), with an estimated population of 528,269 habitants (projection year 2016), a total area of 152.13 km^2 and an urban area of 33.28 km^2, annual mean temperature of 25° and average altitude of 959 m. For the period from January 2009 to December 2015, 25,365 dengue cases (dengue and severe dengue) were geocoded and allocated to 293 census sections and 91 epidemiological periods in Bucaramanga. We employed the official cartography and population data from the 2005 Census in Colombia. A census section (CS) is *"a cartographic-bounded urban division approximately equal to 20 contiguous census blocks and belonging to an urban sector"*, where a census block is a *"lot of land built or unbuilt bounded by vehicular or pedestrian traffic roads of a public nature"* [47].
An epidemiological period (EP) is a component of the epidemiological calendar applied

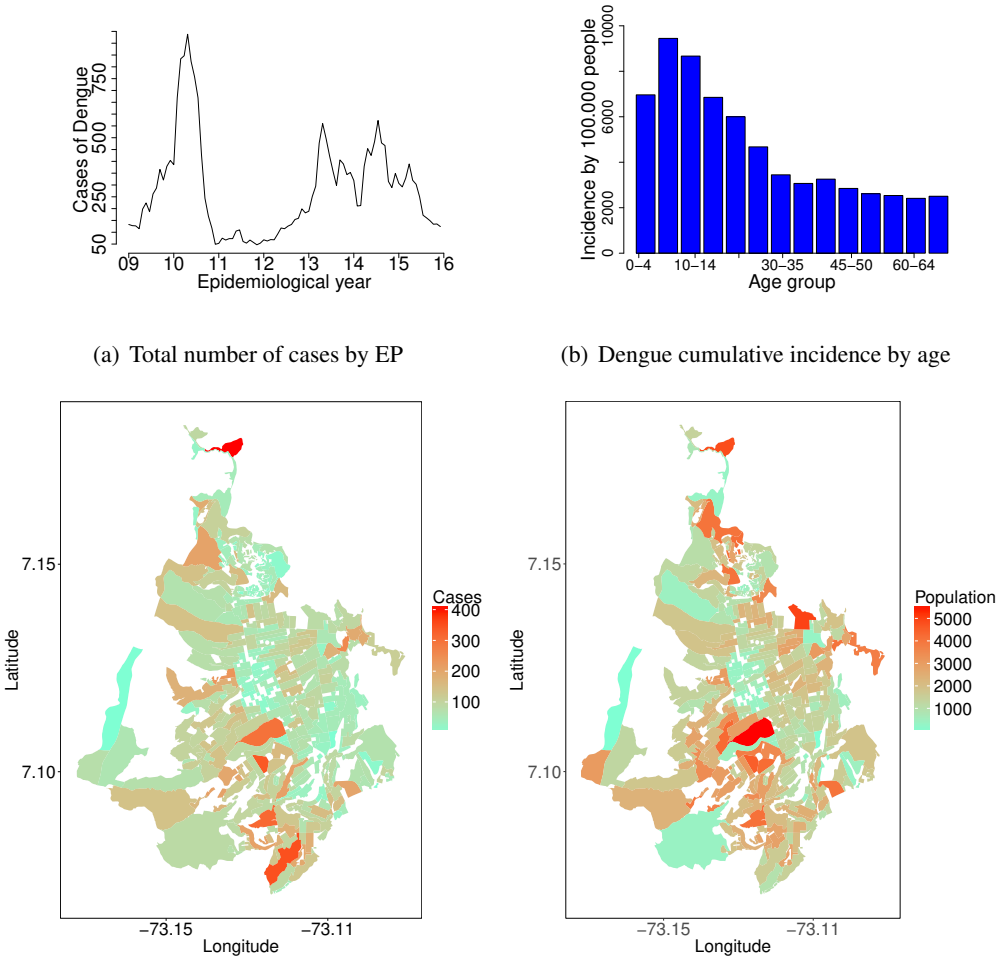
by the Pan-American Health Organization. The epidemiological calendar divides the epidemiological year (EY) in 52 epidemiological weeks (EW), aggregated in 13 EP (one EP equal to four EW). The EW begins on Sunday and ends on Saturday. Our dengue data start at the first EP of 2009 and end at the last EP of 2015, giving a total of 91 EPs (13 EPs times 7 years).
The total dengue incidence rate was calculated by five-years age groups using the city population from 2015 as the population at risk. The expected dengue cases counts by CS and EP were calculated as the product of the dengue incidence rate by age group and the population by age group in every CS (Census 2005), divided by the number of EP.
Figure 4.1(a) displays the total number of cases of dengue by EP. We observed one major outbreak in 2010, and two smaller peaks in incidence in 2013 and 2014. Figure 4.1(b) presents the cumulative dengue disease incidence by five-year age group, showing the highest incidence in people under 20 years of age. The largest CS in terms of population is around 5500 people, and the largest concentrations of people by CS are located in the southwestern and northern areas of the city (Figure 4.1(c)), while the highest number of cases of dengue cases by CS are located in southeastern Bucaramanga (Figure 4.1(d)). Figure 4.2(a) presents the longitudinal profiles of the dengue case counts by CS and EP. The highest number of cases by CS and EP is 21, in mid-2010, followed by 16 cases in 2014. Around 80% of dengue case counts by CS and EP are comprised by zero cases (Figure 4.2(b)). The heatmap of dengue cases by CS and EP (Figure 4.2(c)) shows the epidemic wave in 2010 affecting almost all CS, followed by a period with a low number of cases in 2011 and 2012, the epidemic waves in 2013 and 2014, and the decreasing number of cases in 2015. We can see that most of the CS with a large number of dengue cases in 2010 also display a large number of cases in 2013 and 2014.

4.2.2 Satellite images

We obtained covariates for the modeling process from the Moderate-Resolution Imaging Spectroradiometer (MODIS) satellite images. MODIS is a scientific instrument launched in terrestrial orbit by NASA: in 1999 in the TERRA satellite and in 2002 in the AQUA satellite.

Land surface temperature

We used MODIS MOD11A2 version 6 product [48] eight-day, per-pixel land surface temperature (LST) in a 1200 x 1200 km grid. Each pixel value in the MOD11A2 is a simple average of all the corresponding MOD11A1 LST pixels collected within that eight-day period, and the pixel size is 1000 m. We selected the *day time surface temperature* band.



(a) Total number of cases by EP

(b) Dengue cumulative incidence by age

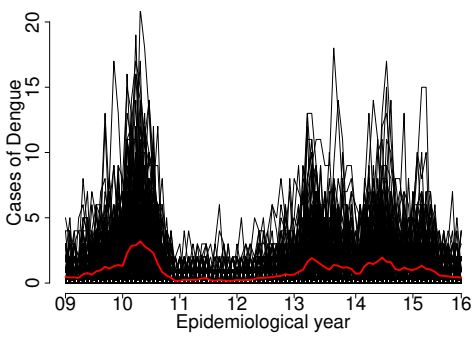
(c) Cases by CS

(d) Population by CS

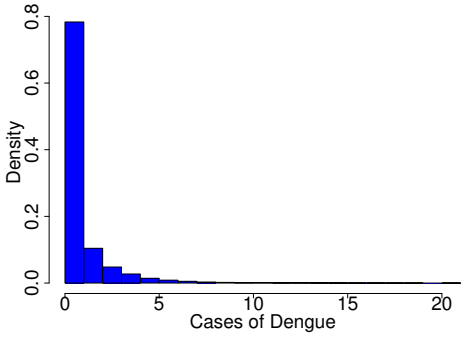
Figure 4.1: Descriptive statistics of cases of dengue disease (a) total number of cases by EP, (b) age-adjusted cumulative incidence by 100,000 people, (c) cases of dengue by CS in Bucaramanga and (d) population by CS, January 2009 - December 2015

Normalized difference vegetation index (NDVI)

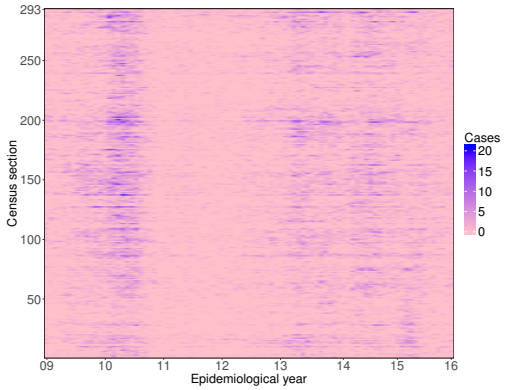
We also employed MODIS MOD13Q1 version 6 product [49] 16-day, vegetation index (VI) value at a per pixel basis. We chose the *normalized difference vegetation index* (NDVI) band, which is referred to as the continuity index by the existing National Oceanic and Atmospheric Administration Advanced Very High Resolution Radiometer (NOAA-AVHRR) derived NDVI. The 16-day composite VI is generated using the two 8-day composite reflectance granules (MxD09A1) in the 16-day period. The grid consists of 4800 rows and 4800 columns of 250 m pixels. The algorithm chooses the best available



(a) Longitudinal plot of dengue by CS and EP



(b) Histogram of dengue cases by CS



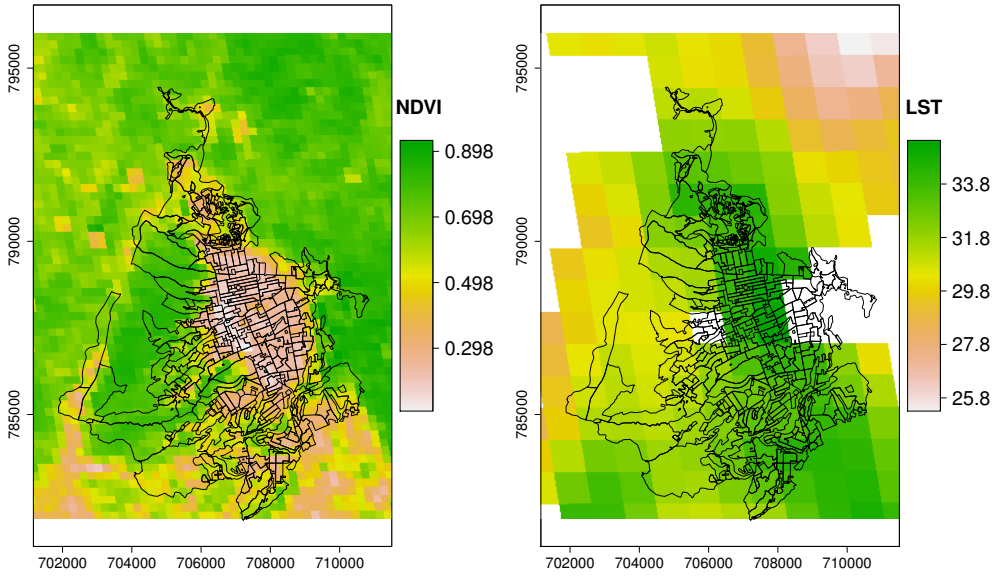
(c) Heatmap of number of dengue cases by CS and EP

Figure 4.2: (a) Longitudinal plots of dengue cases of dengue disease by CS and EP (red line is the average number of cases by EP), (b) histogram of number of dengue cases by CS and EP, and (c) heatmap of dengue cases by CS and EP

pixel value from all the acquisitions from the 16-day period.

4.2.3 Image processing

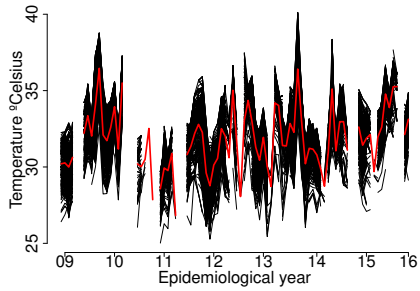
We employed raster images from December 2008 to December 2015, reprojected from Sinusoidal to UTM 18N datum WGS84, and resampled to 30 m using MRT software ([50]). Composite raster images by EP were created by pixel, averaging all the reprojected and resampled images in EP. For every composite raster image by EP, we applied a mask comprised by the polygons (CS) from the vectorial map of the city of Bucaramanga, calculating the average LST or NDVI of all pixels by polygon (CS). Figure 4.3 presents examples of the (a) NDVI raster image and (b) LST raster image, with the polygons of the vector shapefile imposed over the raster images. Figure 4.4(a) presents the longitudinal plots of LST and Figure 4.4(b) the histogram of the LST values. LST was highly volatile



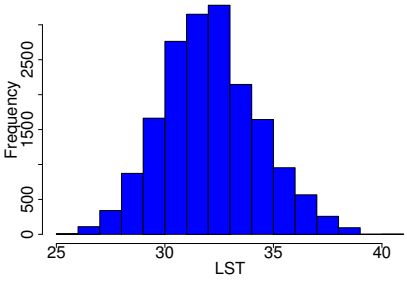
(a) NDVI (b) LST

Figure 4.3: Raster images for (a) NDVI and (b) LST, including the polygons of the vector shapefile map of Bucaramanga. The images correspond to the composite image of the seventh EP, 2009

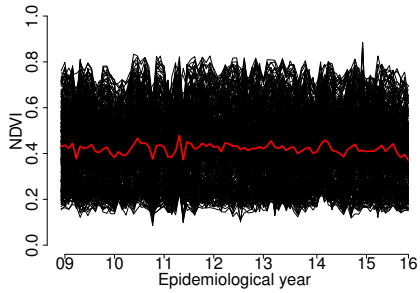
throughout the period, with annual peaks of high LST. There were missing values for some periods, attributed to the clouds over the city. The histogram reveals a peaked and symmetric pattern, with a minimum value around 25°C, a maximum value close to 40°C and average LST close to 32°C. Figure 4.4(c) show the longitudinal profiles of NDVI by CSs, which are fairly stable through the period, displaying a minimum value close to 0.18 units, a maximum around 0.8 units, and an average close to 0.4 units. NDVI presents a two-peaked, fat-tailed density (Figure 4.4(d)). Figure 4.4(e) presents an LST heat-map, revealing the temporal and spatial distribution of the covariate. We can see the missing values noted previously with the second half of 2010 showing the greatest number of missing values. In addition, we observe the clustering of LST between some CSs over time. The NDVI heat map displays no missing values and stability in the NDVI over time (Figure 4.4(f)). We end the exploratory data section presenting the density of correlations by CSs between cases of dengue and NDVI or LST. The correlations between cases of dengue and lag-zero EP and lag-one EP LST are shown in Figure 4.5(a) and Figure 4.5(b), respectively. We found densities slightly skewed to the right, centered around 0.2.
The correlation between cases of dengue and lag-zero EP and lag-one EP NDVI are presented in Figure 4.5(c) and Figure 4.5(d), where the densities are centered around zero.



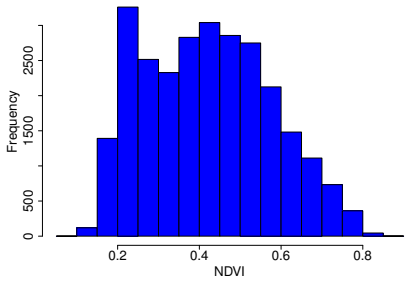
(a) LST longitudinal plot



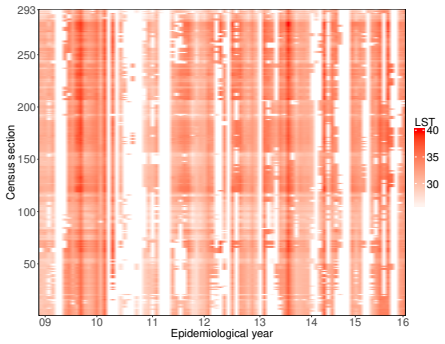
(b) LST histogram



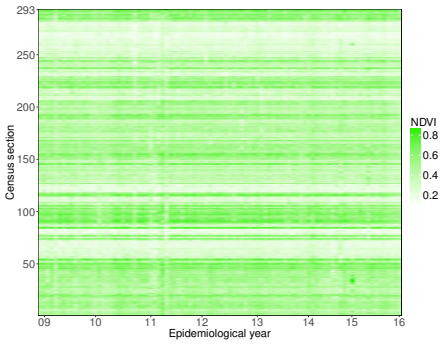
(c) NDVI longitudinal plot



(d) NDVI histogram

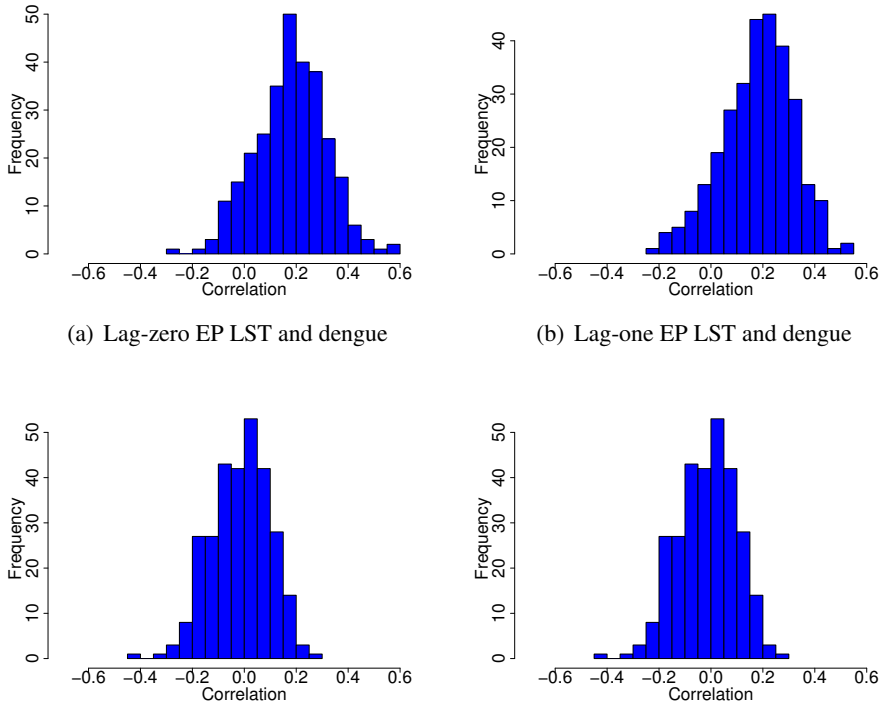


(e) LST heatmap



(f) NDVI heatmap

Figure 4.4: Exploratory graphical analysis of LST and NDVI. (a) Longitudinal plot of LST by CS (red line is the average LST by EP) (b) histogram of LST by CS and EP, (c) longitudinal plots of NDVI by CS (red line is the average NDVI by EP), and (d) histogram of NDVI by CS and EP, and (e) LST and (f) NDVI heat maps by CS and EP



(a) Lag-zero EP LST and dengue

(b) Lag-one EP LST and dengue

(c) Lag-zero EP NDVI and dengue

(d) Lag-one EP NDVI and dengue

Figure 4.5: Histogram of linear correlations between dengue cases and lag-zero EP and lag-one EP NDVI or LST by CS

4.3 Interaction effects models for space-time variation of relative risk

Following Lawson ([12]), for areal (lattice) data the relative risk is estimated when a Poisson model with parameter ω_i is assumed for disease count y_i in the ith small area ($i = 1, \cdots, N =$ total number of small areas).

$$y_i \sim \text{Poisson}(\omega_i) \tag{4.1}$$

The mean function $\text{E}(y_i)$ is defined by two components: the first component is the background population effect calculated by comparing the disease rate in a standard population and a local expected rate. The second component is the relative risk representing the excess risk within a small area. The ratio of observed (y_i) to expected (E_i) counts within small area is called the standardized mortality/morbidity/incidence Ratio (SIR), and this ratio is an estimate of relative risk in the ith area.

$$\text{SIR}_i = \frac{y_i}{E_i} \tag{4.2}$$

SIRs are commonly used in disease mapping, but they have several drawbacks [34]. An improved model-based estimate of relative risk follows the assumption that the data are independently distributed, with expectation

$$\mathrm{E}(y_i) = \omega_i = E_i \exp(\theta_i) \tag{4.3}$$

where E_i is the expected rate, $\exp(\theta_i)$ is the relative risk for the ith small area, and θ_i is a linear predictor with terms for correlated and uncorrelated spatial effects and possibly covariates.
Knorr-Held [43] defines a framework for the space-time variation of risk of areal data, starting with the 'main effects model', where the disease counts y_{it} in the ith area and the tth time period ($t = 1, \cdots, T$ =total number of time periods) are Poisson distributed

$$y_{it} \sim \text{Poisson}(\omega_{it}) \tag{4.4}$$

$$\omega_{it} = E_{it} \exp(\theta_{it}) \tag{4.5}$$

where E_{it} and $\exp(\theta_{it})$ are the expected counts and the relative risk in the ith area and tth time period, and θ_{it} is the linear predictor,

$$\theta_{it} = \alpha + u_i + v_i + \gamma_t + \phi_t \tag{4.6}$$

where the model components are the spatially correlated effects block $\boldsymbol{u}$, the spatially uncorrelated effects block $\boldsymbol{v}$, the temporally correlated effects block $\boldsymbol{\phi}$, and the temporally uncorrelated effects block $\boldsymbol{\gamma}$. The distributional assumptions for each effects block are multivariate Normal with mean zero vector and precision matrix $\tau\mathbf{K}$, where τ is a scalar to be estimated and $\mathbf{K}$ is a known structure matrix.
For the spatially correlated effects block $\boldsymbol{u}$, the precision matrix $\tau_u\mathbf{K}_u$ is a Normal intrinsic conditional autoregression (ICAR) defined according to Besag & Kooperberg ([51]), where the k_{ij} elements of the structure matrix $\mathbf{K}_u$ are -1 if areas i and j are adjacent, 0 if areas i and j are not adjacent or n_i if $i = j$, the n_i is the number of neighbors of the ith area, and τ_u is a precision,

$$\boldsymbol{u} \sim \text{Normal}(\mathbf{0}, \tau_u\mathbf{K}_u) \tag{4.7}$$

The temporally correlated effects $\boldsymbol{\phi}$ are Normal ICAR first order random walk (RW1) with precision matrix $\tau_\phi\mathbf{K}_\phi$, a precision τ_ϕ and a RW1 structure matrix $\mathbf{K}_\phi$,

$$\boldsymbol{\phi} \sim \text{Normal}(\mathbf{0}, \tau_\phi\mathbf{K}_\phi) \tag{4.8}$$

For the spatially and temporally uncorrelated blocks $\boldsymbol{v}$ and $\boldsymbol{\gamma}$, the structure matrices $\mathbf{K}_v$ and $\mathbf{K}_\gamma$ are equal to the identity matrices $\mathbf{I}_v$ and $\mathbf{I}_\gamma$ respectively, determining the spatial uncorrelated effects with precision τ_v,

$$\boldsymbol{v} \sim \text{Normal}(\mathbf{0}, \tau_v\mathbf{I}_v) \tag{4.9}$$

and the temporal uncorrelated effects with precision τ_γ,

$$\boldsymbol{\gamma} \sim \text{Normal}(\mathbf{0}, \tau_\gamma\mathbf{I}_\gamma) \tag{4.10}$$

With the inclusion of ψ_{it} terms to the main effects model, the IE model appears

$$\theta_{it} = \alpha + u_i + v_i + \gamma_t + \phi_t + \psi_{it} \tag{4.11}$$

where the $\boldsymbol{\psi}$ are Normally distributed with zero mean, precision τ_ψ, and structure matrix $\mathbf{K}_\psi$ defined as the Kronecker product of the structure matrices of the spatial and temporal effects,

$$\boldsymbol{\psi} \sim \text{Normal}(0, \tau_\psi \mathbf{K}_\psi) \tag{4.12}$$

then, the structure matrix defines the IE type. First, the Kronecker product between the structure matrices of uncorrelated spatial and temporal effects defines the type I (unstructured) IE model, $\mathbf{K}_\psi = \mathbf{K}_\gamma \otimes \mathbf{K}_v$.
Second, the type II (temporal) IE with structure matrix $\mathbf{K}_\psi$ is defined as the Kronecker product of the structure matrices of the temporally correlated and the spatially uncorrelated effects, i.e $\mathbf{K}_\psi = \mathbf{K}_\phi \otimes \mathbf{K}_v$.
Third, the type III (spatial) IE model is defined by a structure matrix $\mathbf{K}_\psi$ obtained as the Kronecker product of the spatially correlated and the temporally uncorrelated effects structure matrices, i.e. $\mathbf{K}_\psi = \mathbf{K}_u \otimes \mathbf{K}_\gamma$
Finally, the type IV (inseparable) IE model is defined by a structure matrix built as the Kronecker product of the temporally and spatially correlated effects structure matrices, i.e. $\mathbf{K}_\psi = \mathbf{K}_\phi \otimes \mathbf{K}_u$

4.4 Spatiotemporal modeling of relative risk of dengue disease

4.4.1 Modeling relative risk of dengue disease

For modeling the relative risk of dengue disease by the ith CS ($i = 1, \cdots N = 293$) and the tth EP ($t = 1, \cdots T = 91$), we assume the observed counts of dengue disease O_{it} are Poisson-distributed with ω_{it} parameter

$$O_{it} \sim \text{Poisson}(E_{it} \exp\theta_{it}) \tag{4.13}$$

$$\omega_{it} = E_{it} \exp(\theta_{it}) \tag{4.14}$$

where E_{it}, $\exp(\theta_{it})$ and θ_{it} are the expected dengue count, the relative risk of dengue disease and the linear predictor, respectively, by CS i and EP t. We start by fitting the 'main effects model' (Equation 4.6) removing the spatial and temporal uncorrelated effects (v_i and γ_t), and replacing the Normal ICAR priors to the spatially correlated effects u_i by the Leroux conditional autorregressive (CAR) priors ξ_i [52]. The conditional prior probability distribution for the spatially correlated effect ξ_i is defined as

$$\xi_i | \xi_{-i} \sim \text{Normal}\left(\frac{\rho}{(1-\rho+\rho d_i)} \sum_{j \in \partial i} \xi_j, \frac{\sigma_\xi^2}{(1-\rho+\rho d_i)} \right) \tag{4.15}$$

where σ_{ξ}^2 is the variance of the ξ_i, d_i is the number of neighbors for the ith CS, $j \in \partial_i$ are the jth neighbors of the ith CS, and ρ is a smoothing parameter, determining the degree of association between the spatially correlated effects. ρ close to 1 indicates a predominance of spatial correlation, and ρ close to 0 is associated with spatially uncorrelated effects. For the intercept α we assign an improper Uniform prior.
Normal ICAR RW1 priors with precision τ_ϕ are assigned to the temporally correlated effects ϕ_t. The model formulation ends assigning Uniform(0,1) hyperpriors for $\tau_\phi^{-1/2}, \rho,$ and σ_ξ. Then, the linear predictor for the main effects model is

$$\theta_{it} = \alpha + \xi_i + \phi_t \tag{4.16}$$

Next, we fitted type I to III IE models, adding the interaction terms ψ_{it}

$$\theta_{it} = \alpha + \xi_i + \phi_t + \psi_{it} \tag{4.17}$$

We start with the type I (unstructured) IE model, where the $\boldsymbol{\psi}$ are Normal with mean zero vector, precision τ_ψ with Uniform(0,1) hyper-prior to $\tau_\psi^{-1/2}$, and structure matrix $\mathbf{K}_\psi = \mathbf{I}_v \otimes \mathbf{I}_\gamma$, and $\boldsymbol{\psi}$ defined in Equation 4.12.
We then fitted type II (temporal) IE models where the $\boldsymbol{\psi}$ are Normal ICAR RW1 with global (τ_ψ) or local precision (τ_{ψ_i}). The type II IE model with the global precision (τ_ψ) specifies a single precision for all the temporal IE, with Uniform(0,1) hyperprior for $\tau_\psi^{-1/2}$, and structure matrix $\mathbf{K}_\psi = \mathbf{K}_\phi \otimes \mathrm{I}_v$, determining the $\boldsymbol{\psi}$ presented in Equation 4.12.
The type II IE model with local precision defined by the precision vector $\boldsymbol{\tau}_{\psi_{LT}} = (\tau_{\psi_1}, \cdots, \tau_{\psi_N})$ with Uniform(0,1) hyper-priors for $\tau_{\psi_i}^{-1/2}$. Then, the precision matrix $\mathbf{Q}_{\psi_T} = \text{Diagonal}(\boldsymbol{\tau}_{\psi_{LT}}) \otimes \mathbf{K}_\phi$ establishes the probability distribution

$$\boldsymbol{\psi} \sim \text{Normal}(\mathbf{0}, \mathbf{Q}_{\psi_T})$$

Then, we fitted type III (spatial) IE models where the $\boldsymbol{\psi}$ are Normal ICAR with global (τ_ψ) or local precision (τ_{ψ_t}). The type III IE model with global precision τ_ψ implies one precision for all IE, receiving a Uniform(0,1) hyper-prior for the $\tau_\psi^{-1/2}$, structure matrix $\mathbf{K}_\psi = \mathbf{K}_u \otimes \mathrm{I}_\gamma$, and the $\boldsymbol{\psi}$ with probability distribution presented in Equation 4.12.
The type III IE model with local precision vector $\boldsymbol{\tau}_{\psi_{LS}} = (\tau_{\psi_1}, \cdots, \tau_{\psi_T})$ with Uniform(0,1) hyperpriors for $\tau_{\psi_t}^{-1/2}$, precision matrix $\mathbf{Q}_{\psi_S} = \text{Diagonal}(\boldsymbol{\tau}_{\psi_{LS}}) \otimes \mathbf{K}_u$, and probability distribution

$$\boldsymbol{\psi} \sim \text{Normal}(\mathbf{0}, \mathbf{Q}_{\psi_S}) \tag{4.18}$$

Finally, we fitted type IV (inseparable space-time) IE models, dropping the temporally correlated effect ϕ_t

$$\theta_{it} = \alpha + \xi_i + \psi_{it} \tag{4.19}$$

We followed the implementation from Lagazzio *et al.* [53], where the IE ψ_{it} are normally distributed

$$\psi_{it} | \psi_{-it} \sim \text{Normal}(\mu_{\psi_{it}}, \tau_\psi d_{it}) \tag{4.20}$$

where $\mu_{\psi_{it}}$ is the mean of the conditional distribution of ψ_{it} given all the other ψ, and $\tau_\psi d_{it}$ is a weighted precision obtained as a product of an overall precision τ_ψ and a Uniform(0,2) hyperprior for $\tau_\psi^{-1/2}$, and the weight d_{it} dictated by $d_{it} = d_i$, for $t = 1$ or $t = T$, or $d_{it} = 2d_i$ otherwise. The formulation ends with $\mu_{\psi_{it}}$, defined as

$$\mu_{\psi_{it}} = \begin{cases} \psi_{i2} + \sum_{j\in\partial_i} \frac{\psi_{j1}}{d_i} - \sum_{j\in\partial_i} \frac{\psi_{j2}}{d_i} & t = 1 \\ \frac{1}{2}(\psi_{i,t-1} + \psi_{i,t+1}) + \sum_{j\in\partial_i} \frac{\psi_{jt}}{d_i} - & \\ \sum_{j\in\partial_i} \frac{\psi_{j,t+1}+\psi_{j,t-1}}{2d_i} & t = 2,\cdots,T-1 \\ \psi_{i,T-1} + \sum_{j\in\partial_i} \frac{\psi_{jT}}{d_i} - \sum_{j\in\partial_i} \frac{\psi_{j,T-1}}{d_i} & t = T \end{cases}$$

4.4.2 Modeling LST missing values

For the imputation of LST missing values, we assumed the LST_{it} are Normally distributed with mean μ_{it} and precision τ_{LST}.

$$\text{LST}_{it} \sim \text{Normal}(\mu_{it}, \tau_{\text{LST}}) \tag{4.21}$$

where μ_{it} is a linear predictor, and a Gamma(2,0.01) hyperprior is assigned to the precision τ_{LST}. This Gamma hyperprior has a mean of 200 and a variance of 20000, determining a noninformative prior for the precision. We assumed a type II (temporal) IE model with local precision. The linear predictor follows,

$$\mu_{it} = \alpha + u_i + \phi_t + \psi_{it} \tag{4.22}$$

where the spatially correlated effects $\boldsymbol{u}$ are Normal ICAR with mean zero vector, precision τ_u and structure matrix $\mathbf{K}_u$, and probability distribution presented in Equation 4.7.
The temporally correlated effects $\boldsymbol{\phi}$ are Normal ICAR RW1 with mean zero vector, precision τ_ϕ and structure matrix $\mathbf{K}_\phi$, with probability distribution displayed in Equation 4.8.
The temporal IE $\boldsymbol{\psi}$ are Normal ICAR RW1 with mean zero vector, local precision vector $\boldsymbol{\tau}'_{\psi_{local}} = (\tau_{\psi_1}, \cdots, \tau_{\psi_N})$, and precision matrix $\mathbf{Q}_\psi = \text{Diagonal}(\boldsymbol{\tau}_{\psi_{local}}) \otimes \mathbf{K}_\phi$

$$\boldsymbol{\psi} \sim \text{Normal}(\mathbf{0}, \mathbf{Q}_\psi) \tag{4.23}$$

The model is completed with Gamma(0.5,0.005) (mean of 100 and variance of 20000) non-informative hyperpriors for τ_{ψ_i}, τ_ϕ and τ_u. The type IV IE model was applied to impute the missing LST data under the assumptions of missing completely at random as missingness mechanism, and, the strong association of LST in the same CS throughout the time periods.

4.4.3 Spatiotemporal interaction effects models including covariates

The modeling process involved the following stages: first we fitted the IE models without covariates presented in Section 4.4.1. Second, we imputed the LST values using the model in Section 4.4.2. With the complete data for LST and the NDVI data, we fitted the

IE model with covariates. We employed lag-zero EP or lag-one EP covariates (LST and NDVI) included in two forms: fixed coefficients and time-varying coefficients (TVC). The use of lag-zero and lag-one EP covariates is justified by field research in dengue, which shows probable dengue cases up to three weeks prior or subsequent to the time point of interest along with negligible evidence of cyclical or seasonal patterns in dengue cases by CS. The full IE model with fixed coefficients for the lag-zero EP covariates follows,

$$\theta_{it} = \alpha + \xi_i + \phi_t + \psi_{it} + \beta_1 \text{NDVI}_{i,t} + \beta_2 \text{LST}_{i,t}$$

and the lag-one EP covariates

$$\theta_{it} = \alpha + \xi_i + \phi_t + \psi_{it} + \beta_1 \text{NDVI}_{i,t-1} + \beta_2 \text{LST}_{i,t-1}$$

where the fixed coefficients β_k ($k = 1$ For NDVI, $k = 2$ for LST) are Normally distributed with mean zero and precision 0.01 (variance of 100), configuring a non-informative prior for the fixed coefficients,

$$\beta_k \sim \text{Normal}(0, 0.01)$$

Next, we fitted type I to IV IE models with auto-regressive TVC $b_{k,t}$ to the lag-zero EP covariates

$$\theta_{it} = \alpha + \xi_i + \phi_t + \psi_{it} + b_{1,t} \text{NDVI}_{it} + b_{2,t} \text{LST}_{it}$$

and lag-one EP covariates,

$$\theta_{it} = \alpha + \xi_i + \phi_t + \psi_{it} + b_{1,t} \text{NDVI}_{i,t-1} + b_{2,t} \text{LST}_{i,t-1}$$

where the coefficient for $t = 1$ is Normal with mean zero and precision 0.1 (variance of 10)

$$b_{k,1} \sim \text{Normal}(0, 0.1)$$

and the coefficients for $t = 2, \cdots, T$ are Normal with mean $b_{k,t-1}$ and precision 10 (variance of 0.1)

$$b_{k,t} \sim \text{Normal}(b_{k,t-1}, 10)$$

The choice of priors follows Lee and Shaddick [54] and Martínez-Bello *et al.* [55], allowing the TVC at $t = 1$ a wide space for the parameter search, and for $t = 2, \cdots, T$ restricting the parameter space assigning a small variance with the objective of smoothing the TVCs trend throughout the study period.

4.4.4 Inference and model selection

For parameter estimation we applied Markov Chain Montecarlo (MCMC) simulations, using WinBUGS 1.4 [56] and OpenBUGS 3.2.3. [57]. Models ran for 60000 iterations total, with a 50000 iterations burn-in, thinning of 10, and 2 chains, producing 2 chains of 1000 iterations for inferences. The main effects model and the type I to III IE models were fitted in WinBUGS, and the type IV IE models were fitted in OpenBUGS. We assessed chains convergence using trace and density plots of selected parameters as well as the Gelman and Rubin statistic [58]. For model selection, we recovered the log predictive posterior simulations for each data point and calculated the deviance, the widely acceptable information criterion (WAIC) [59], leave-one-out cross-validation (LOO) [60] and the effective number of parameter (p_{WAIC} and p_{LOO}) using the R package `loo` [61].

4.5 Results of the inference

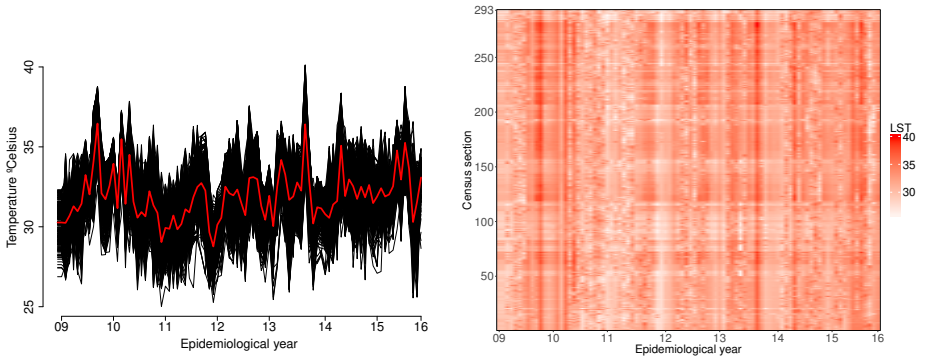
In this section, we first present the results for the selection process of the IE models without covariates, then the results from the selected IE model plus covariates, and finally the model parameters from the selected final model. Table 4.1 presents the deviance, LOO and WAIC for the models without covariates. The two models with the lowest deviance were the type IV IE model (deviance = 54197.1) and type I IE model (deviance = 54264.9). The two models with the lowest LOO and WAIC were the type IV IE model (LOO = 58623.5, WAIC = 58270.2) and the type II IE model with local precision (LOO = 58737.5, WAIC = 58528.8). At this stage we selected the type IV IE model as the base model to include covariates with fixed and time-varying coefficients. Then, we imputed the LST

Table 4.1: Deviance, leave-one-out cross-validation (LOO) and WAIC for spatiotemporal interaction effects models without covariates, for the relative risk of dengue disease, January 2009 - December 2015

Interaction	Deviance	LOO	p_{LOO}	WAIC	p_{WAIC}
Main effects	60915.7	61419.5	490.4	61418.5	489.9
I	54264.9	59994.3	4468.6	59204.7	4073.8
II, global	56426.0	59112.9	2378.8	59028.0	2336.3
II, local	55394.7	58737.5	2857.5	58528.8	2753.1
III, global	55519.7	59572.9	3340.4	59221.3	3164.6
III, local	55614.0	59543.3	3243.1	59200.3	3071.6
IV	54197.1	58623.5	3652.3	58270.2	3475.6

missing values through the type II IE model with local precision defined in Section 4.4.2. The complete LST imputed dataset appears in Figure 4.6, where panel (a) shows the longitudinal plots, and panel (b) presents the heatmap of LST by CS and EP. The type II IE model imposes an structure to the imputed LST, where the LST in every CS is highly

related to the LST in the previous and the following EP in the same CS, with overall spatially and temporally correlated effects throughout all the EPs. Table 4.2 presents the



(a) Imputed LST longitudinal plot (b) Imputed LST Heatmap

Figure 4.6: LST imputed mean values from the interaction type II model with local precision: (a) longitudinal plot (red line is the LST average by EP), and (b) LST heatmap by CS and EP

results for the type IV IE models with lag-zero and -one EP covariates. The two models with the lowest deviance are the model including time-varying coefficients for lag-one EP NDVI (deviance = 54184.9) and the model with constant coefficient for lag-one EP NDVI (deviance = 54179.2). The two models with the lowest LOO cross-validation are the model with constant coefficient for lag-zero EP LST (LOO = 58621.9) and the model with constant coefficients for lag-one EP NDVI plus LST (LOO = 58627.0), while the models with the lowest WAIC are the model with constant coefficient for lag-zero EP LST (WAIC = 58267.5) and the model with constant coefficient for lag-one EP NDVI (WAIC = 58276.2). Here, we choose for inference the type IV IE model plus constant coefficient for lag-zero EP LST, showing the model parameters in Table 4.3. The model parameters reveal higher variability for α, ρ and σ_u than for the σ_ψ. The parameter ρ, which denotes the importance of the spatially correlated effects, is centered at 0.6319, suggesting medium to high importance for the spatial effects as determinants of high risk of dengue disease in certain CSs. The constant coefficient for the lag-zero EP LST presents a positive but weak association (95% credible intervals including zero) between LST and cases of dengue.

Figure 4.7(a) shows that the temporal trend (which is an overall measure of the risk evolution through the study period) is above zero for the first half of 2010, denoting the period with the highest risk. It is under zero for the years 2009, 2011, 2012 and the end of 2015, and close to zero for the years 2013 and 2014, showing that the overall risk for those two years was average with respect to the overall period.

Figure 4.7(b) displays the map of the spatial effects ξ_i, showing a cluster of high spatial effects concentrated around the CSs in the southeastern part of the city. Figure 4.8 shows heat maps for the logarithm of the mean relative risk (Log RR) and the logarithm of the standardized incidence rate (Log SIR) for dengue disease by CS and study period. We

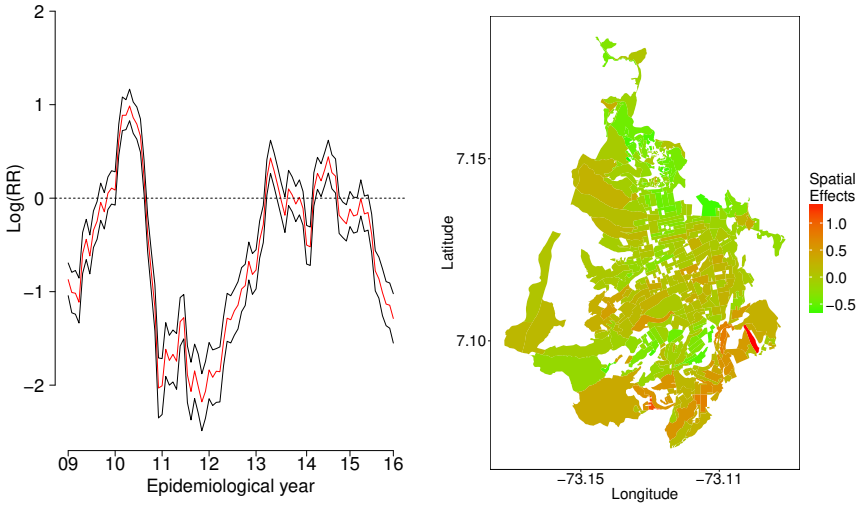
Table 4.2: Deviance, LOO cross-validation and WAIC for the type IV IE models plus lag-one or lag-zero covariates for the relative risk of dengue, Jan. 2009 - Dec. 2015

Model	Deviance	LOO	p_{LOO}	WAIC	p_{WAIC}
No covariates	54197.1	58623.5	3652.3	58270.2	3475.6
Lag-zero EP covariates					
$NDVI_t$	54191.1	58650.2	3675.2	58287.0	3493.6
LST_t	54261.8	58725.6	3681.0	58363.8	3500.1
NDVI	54187.3	58640.9	3670.7	58278.1	3489.3
LST	54210.4	58621.9	3640.6	58267.5	3463.4
NDVI + LST	54261.1	58631.4	3613.1	58288.5	3441.7
Lag-one EP covariates					
$NDVI_t$	54184.9	58645.1	3673.1	58276.3	3488.7
LST_t	54209.6	58698.8	3697.5	58332.5	3514.4
NDVI	54179.2	58642.1	3676.2	58276.2	3493.2
LST	54236.0	58631.9	3630.5	58283.5	3456.3
NDVI + LST	54241.1	58627.0	3625.1	58283.9	3453.5

Table 4.3: Parameter estimates for the type IV IE model plus fixed coefficient for lag-zero EP LST, Jan. 2009 - Dec. 2015

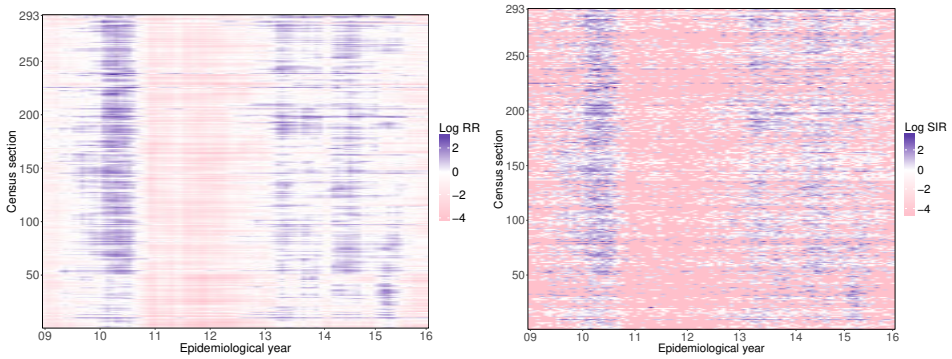
Parameter	Mean	SD	Q2.5%	Q50%	Q97.5%
α	-0.8713	0.0943	-1.0430	-0.8739	-0.6931
β_{LST}	0.0001	0.0068	-0.0315	0.0001	0.0315
ρ	0.6319	0.2342	0.1457	0.6581	0.9814
σ_ψ	0.5543	0.0608	0.4603	0.5612	0.6423
σ_u	0.6500	0.1215	0.4296	0.6525	0.8897

observed the smoothing effect of the modeling process, which allows the visualization of general patterns of the risk distribution in time and space. Finally, we present the key output from the spatiotemporal modeling of the relative risk of dengue disease. Figure 4.9(a) shows the logarithm of the mean relative risk of dengue disease, and Figure 4.9(b) the discretized lower bound (DLB) of the 95% credible interval (CI) of the logarithm of relative risk, between the sixth and the eleventh EP of 2010, a period including one major outbreak in the city. The maps let us explore the evolution of the relative risk between CSs over time. In Figure 4.9(a), we distinguish some areas in the southwest, northwest and east presenting high risk from EP 6 to EP 9, before returning to low risk at the EPs 10 and 11 of 2010. The maps of the DLB 95% CI of the logarithm of relative risk in Figure 4.9(b) show the CS with the relative risk logarithm greater than 0 with 95% probability, facilitating the detection of areas at high risk of dengue. We



(a) (b)

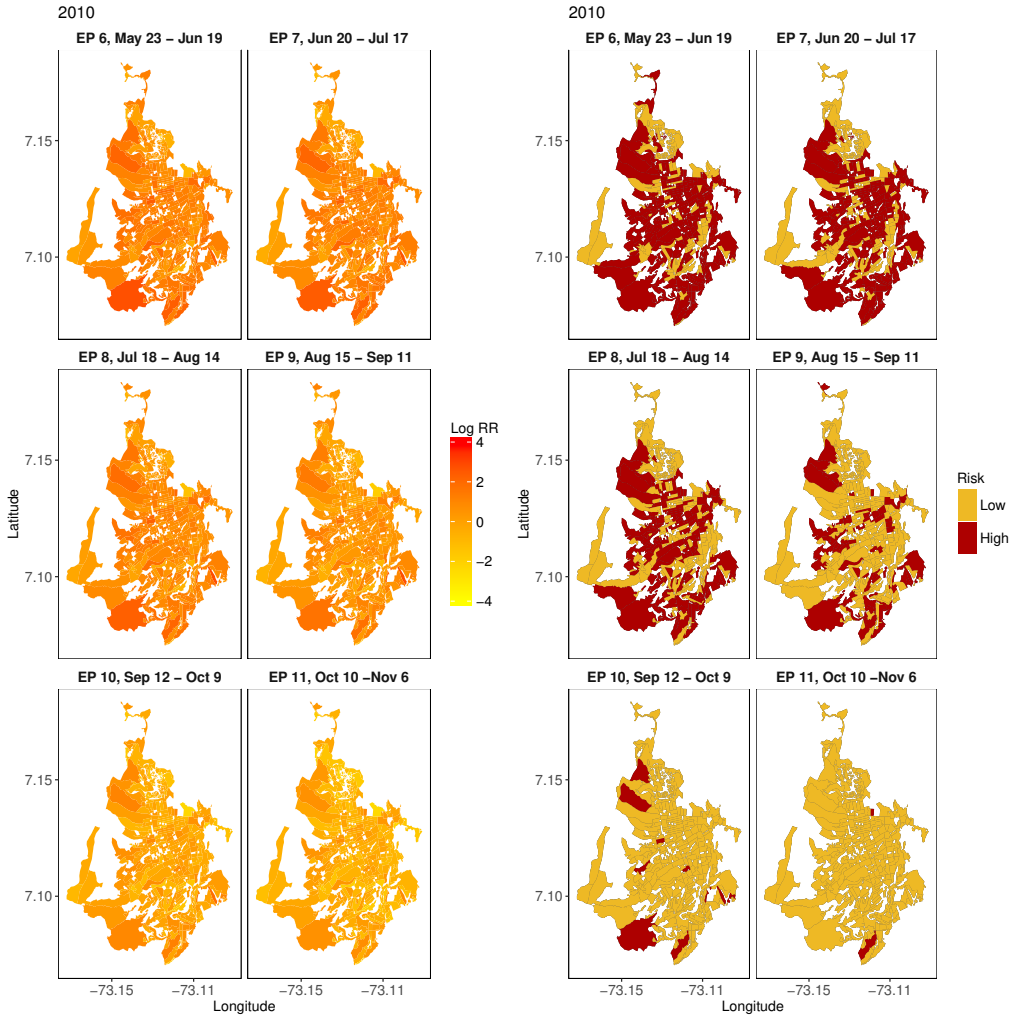
Figure 4.7: (a) Temporal trend of risk, and (b) spatial effects u_i from type IV IE model plus constant coefficient for lag-zero EP LST



(a) Log RR (b) Log SIR

Figure 4.8: Logarithm of the mean relative risk (Log RR) and logarithm of the standardized incidence rate (Log SIR) of dengue disease by CS and EP

have included online supplementary material consisting of the maps of the logarithm of mean relative risk and the DLB 95% CI of the logarithm of relative risk for all the epidemiologic periods of 2009-2015, processed from the output of the spatiotemporal type IV IE model including fixed coefficient of LST.



(a) Log RR

(b) DLB 95% CI Log RR

Figure 4.9: (a) Logarithm of the mean relative risk of dengue disease in Bucaramanga by CS from the sixth EP until the eleventh EP of 2010; (b) discretized lower bound of the 95% credible interval of relative risk of dengue disease (DLB 95% CI Log RR). DLB 95% CI Log RR ≤ 0: low risk; DLB 95% CI Log RR > 0: high risk.

4.6 Discussion

This paper presents spatiotemporal models for relative risk of dengue disease plus covariates obtained from satellite images in a Colombian city over 91 EPs and across 293 CSs. The modeling process reveals the best model to be the type IV (inseparable) IE model including constant coefficient for the lag-zero EP LST, based on the analysis of the WAIC and LOO information criteria. The conclusion from this selection is that the relative risk of dengue disease in every CS depends not only on the relative risk of the neighboring CS, but also on the relative risk from the same CS in the previous and subsequent period as well as the relative risk from neighboring sections in the previous and the subsequent time periods. The interpretation is that the CSs with high relative risk are surrounded by high risk neighboring sections in previous and subsequent periods. The model includes a constant coefficient for the lag-zero EP LST. This choice is based on a consideration of the average LST by CS and EP as an indicator of environmental temperature, in a urban environment.

The first important contribution that this study provides to dengue research consists of its application of spatiotemporal models of relative risk of dengue disease, which have not been previously reported in the literature. Recent systematic reviews have presented studies showing modeling tools for dengue risk mapping [2] and the impact of climate on dengue transmission [62], but none of them presented spatiotemporal models of relative risk of dengue as shown in the manuscript. Lowe *et al.* [35, 39, 40, 42] and Stewart-Ibarra [24] employed Bayesian Poisson and Negative Binomial spatiotemporal relative risk models for early warning systems of dengue disease in Brazil, Thailand and Ecuador; however, those studies only applied main effects models following the framework of Knorr-Held [43], so they did not fully explore models including spatiotemporal interaction-effects. Wijayanti *et al.* [38] fitted dengue and climate data using the type I (unstructured) interaction effects model, but they did not explore the type II to IV interaction effects spatiotemporal models of relative risk.

Another important contribution to the field is our use of spatiotemporal relative risk models to study the association of environmental data obtained from satellite sensors and dengue disease using areal (lattice) data. The final best model included a fixed coefficient of lag-zero EP LST and type IV (inseparable) interaction effects. Several studies associated LST and dengue data with diverse findings. Nazri *et al.* [63] studied the association between risk of dengue disease with covariates obtained from Landsat (LST and NDVI) in the district of Subang-Jaya, Malaysia, using commercial GIS tools. They did not find any association between NDVI and increased risk of dengue but they found a positive correlation between the LST in surrounding lowlands and high risk of dengue disease by district (we obtained a weak association between LST and dengue by census section). In the urban setting of Punta Arenas, Costa Rica, dengue fever has been directly related to vegetation and inversely associated with built environment at a local scale within the urban environment, when those characteristics were assessed using satellite imagery from MODIS (Troyo *et al.* [64]). Buczak *et al.* [65] related meteorological, demographical and satellite (rainfall, NDVI, LST) variables obtained at several resolution to dengue

cases in Peru using an expert system for predicting new cases. Araujo *et al.* ([66]) used data on thermal remote sensing and GIS to identify urban heat islands in São Paulo and study the influence of environmental and socioeconomic factors potentially associated with mosquito breeding and the incidence of dengue. They found low vegetation cover areas presented higher LST and higher incidence rates, a finding similar to our results. Particularly in Colombia, Poveda *et al.* [67] studied the relationship between satellite imagery of vegetation and malaria incidence at national, regional and municipal spatial scales and at annual, quarterly and monthly time scales, and Quintero *et al.* [68] related information from satellite images for rainfall and LST, meteorological data from land stations and dengue disease at regional level in the Risaralda Province, Colombia with a methodology based on Poisson multiple regression model; however, they did not report an association between dengue cases and LST.

All those studies are relevant to our research in the sense that they relate environmental and meteorological variables from different sources to dengue disease using temporal or spatiotemporal analysis, but they differ from the setting shown in this paper in that they use much higher spatial scales than we did as we modeled the relative risk between the areas at CS, with short aggregation time periods. We include data from satellite images in the risk model, trying to gather higher-resolution data (CS in city) compared with settings at regional or national scale. Remote sensing images associated with epidemiological studies of human diseases initially promised high returns, leading in some cases to frustration due to the difference between satellite data providers, the gap between the resolution scale provided by the image and the required information detail at land scale, and many other factors, but mainly due to the need for holistic frameworks where the remote sensing information could fit as a tool for risk assessment and outbreak management [69].

However, for malaria, Kazansky *et al.* ([70]) described the current knowledge of satellite-based environmental data to predict disease risk and study the challenges and oportunities for including this data in a malaria early warning systems, defining the components of a systems architecture framework for the early warning system, and highlighting the need for effective coordination and collaboration across public and private sector organizations.

In addition, this paper makes an important contribution to the Colombian context which has a high incidence of dengue disease. Disease mapping of dengue disease at small spatial scale is not currently used as an epidemiological tool in Colombia. Relative risk maps of dengue disease could be produced at small spatial (census block, census section or census sector) and temporal scales (epidemiological week, epidemiological periods, or different aggregation levels of epidemiological weeks), supporting public health planning, decision making and intervention. The mapping could be done aggregating the data from a selected time period, estimating expected counts, and applying spatial relative risk models to elaborate the map, or it could be done as we propose in the paper, following the spatiotemporal nature of the data, estimating expected counts by area and time period and then applying main effects models in a first approach, followed by the fitting of spatiotemporal interaction-effects models.

An important effort is required to implement disease mapping techniques for dengue: to create a geocoding team to geocode the cases and the people at risk from the health registers, an epidemiological and biostatistical team to develop the modeling process,

and then a geographical information systems team to make the generated spatiotemporal information accessible to decision makers and field researchers.
Our study also has some limitations. First, we consider that there was some under-reporting, as some people suffering from dengue disease did not see a physician. In addition, there are still shortcomings in the notification system. Romero-Vega *et al.* [71] describe the notification coverage difficulties for dengue disease in Colombia. Second, we employ notification data produced by the physicians. In Colombia, dengue is subject to obligatory notification. Physicians report the case to the online surveillance network based principally on clinical symptoms, and the diagnosis is not always confirmed using laboratory methods. In addition, the dengue case-address data sometimes are not well registered, leading to difficulties in the geocoding process. Third, our methods make a strong assumption that risk of dengue is related to the people's CS of residence and ignores the possibility that people may be infected elsewhere, for example, children who are infected at school, or adults infected at the workplace. Fourth, our study lacks information on vector density, species, and vector biology. These data are scarce or null for the city of Bucaramanga. Mapping disease risk in Bucaramanga could set the basis for synchronizing vector surveillance with dengue reports improving control and prevention of dengue disease through integrated approaches in research and technology transfer for surveillance and control of arboviral diseases.
Even considering these limitations, our study is important in the sense that it uses disease risk mapping techniques that have been established in the statistical literature, but which are still not been employed as a concomitant tool for the control of dengue and many other diseases in urban settings in Colombia. We are aware of the recent explosion in the development of estimation methods, disease mapping techniques, web integration and software (WinBUGS, OpenBUGS, INLA, BayesX, etc.), but their full potential will not be realized until they are applied to solve public health problems like dengue disease.
For the epidemiological study of dengue disease, our research is a contribution to the global effort to understand the determinants of disease in space and time. In the future, we plan to include data from other sources with small resolution in time and space, for covariates like rainfall and humidity; to integrate spatiotemporal relative risk mapping of dengue disease to on-line platforms as shown by Chen *et al.* [20]; and to fit spatiotemporal relative risk models with more than one level of aggregation as presented by Ugarte *et al.* [72].

4.7 References

[1] WHO (2012) *Global Strategy for dengue Prevention and Control, 2010-2020*. World Health Organization. Geneva, Switzerland.

[2] Louis VR, Phalkey R, Horstick O, Ratanawong P, Wilder-Smith A, Tozan Y, Dambach P. Modeling tools for dengue risk mapping - a systematic review. *International Journal of Health Geography.* 2014; **13**(1):50.

[3] Knox, G. The detection of space-time interactions. *Applied Statistics.* 1964; **13**:25-29.

[4] Moran, PAP. Notes on Continuous Stochastic Phenomena. *Biometrika.* 1950; **37** (1): 17–23.

[5] Getis A, Ord JK. The Analysis of Spatial Association by Use of Distance Statistics. *Geographical Analysis.* 1992; **24**(3)

[6] Anselin L. Local Indicators of Spatial Association - LISA. *Geographical Analysis.* 2015; **27**(2): 93-115.

[7] Kulldorff M, Heffernan R, Hartman J, Assunção RM, Mostashari F. A space-time permutation scan statistic for the early detection of disease outbreaks. *PLoS Medicine.* 2005; **2**: 216-224.

[8] Banerjee S, Carlin BP, Gelfand AE. Hierarchical Modeling and Analysisi for Spatial Data, Second Edition. CRC Press. Taylor and Francis Group. Boca Raton, Fl. 2005.

[9] Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning. Data mining, inference and prediction, second edition*. Springer-Verlag New York. 2012 .

[10] Christakos G. *Modern spatiotemporal geostatistics.* Oxford University Press, New York. 2016.

[11] Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data: A Bayesian perspective.. *Statistica Sinica.* 2004; **14**: 731–761.

[12] Lawson A. *Bayesian disease mapping : hierarchical modeling in spatial epidemiology.* 2009. Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton, FL.

[13] Estallo EL, Carbajo AE, Grech MG, Frías-Céspedes M, López L, Lanfri MA, Ludueña-Almeida FF, Almirón WR. Spatio-temporal dynamics of dengue 2009 outbreak in Córdoba City, Argentina. *Acta Tropica.* 2014; **136**(1):129–136.

[14] Morato DG, Barreto FR, Braga JU, Natividade MS, da Costa MCN, Morato V, Da Teixeira MGLC. The spatiotemporal trajectory of a dengue epidemic in a medium-sized city. *Memorias do Instituto Oswaldo Cruz.* 2015; **110**(4):528–533,

[15] Toan DTT, Hu W, Quang Thai P, Hoat LN, Wright P, Martens P. Hot spot detection and spatio-temporal dispersion of dengue fever in Hanoi, Vietnam. *Global Health Action.* 2013; **6**:18,632.

[16] Alzahrani AG, Al Mazroa MA, Alrabeah AM, Ibrahim AM, Mokdad AH, Memish ZA Geographical distribution and spatio-temporal patterns of dengue cases in Jeddah governorate from 2006-2008. *Transactions of the Royal Society of Tropical Medicine and Hygiene.* 2013; **107**(1):23–29,

[17] Ling CY, Gruebner O, Krämer A, Lakes T. Spatio-temporal patterns of dengue in Malaysia: Combining address and sub-district level. *Geospatial Health.* 2014; **9**(1):131–140.

[18] Anno S, Imaoka K, Tadono T, Igarashi T, Sivaganesh S, Kannathasan S, Kumaran V, Surendran SN.Space-time clustering characteristics of dengue based on ecological, socio-economic and demographic factors in northern Sri Lanka. *Geospatial Health.* 2015; **10**(2):376.

[19] Acharya BK, Cao C, Lakes T, Chen W, Naeem S (2016) Spatiotemporal analysis of dengue fever in Nepal from 2010 to 2014. *BMC Public Health*. 2010; **16**(1):849,

[20] Chen CC, Teng YC, Lin BC, Fan IC, Chan TC.Online platform for applying space–time scan statistics for prospectively detecting emerging hot spots of dengue fever. *International Journal of Health Geographics.* 2016; **15**(1):43.

[21] Gil JF, Palacios M, Krolewiecki AJ, Cortada P, Flores R, Jaime C, Arias L, Villalpando C, Alberti DÁmato AM, Nasser JR, Aparicio JP. Spatial spread of dengue in a non-endemic tropical city in northern Argentina. *Acta Tropica.* 2016; **158**:24–31.

[22] Mondini A, Chiaravalloti-Neto F. Spatial correlation of incidence of dengue with socioeconomic, demographic and environmental variables in a Brazilian city. *Science of the Total Environment.* 2008; **393**(2-3):241–248,

[23] Lin CH, Wen TH, Teng HJ, Chang NT. The spatio-temporal characteristics of potential dengue risk assessed by Aedes aegypti and Aedes albopictus in high-epidemic areas. *Stochastic Environmental Research and Risk Assessment.* 2016; **30**(8):2057–2066.

[24] Stewart-Ibarra AM, Muñoz ÁG, Ryan SJ, Ayala EB, Borbor-Cordova MJ, Finkelstein JL, Mejía R, Ordoñez T, Recalde-Coronel GC, Rivero K. Spatiotemporal clustering, climate periodicity, and social-ecological risk factors for dengue during an outbreak in Machala, Ecuador, in 2010. *BMC Infectious Diseases.* 2014; **14**(1):610.

[25] Marzzano de Carvalho R, Nascimento LFC. Space-time description of dengue outbreaks in Cruzeiro, São Paulo, in 2006 and 2011. *Revista da Associação Médica Brasileira.* 2014; **60**(6):565–570,

[26] Xavier DR, Figueiredo Mafra Magalhães MA, Gracie R, dos Reis IC, de Matos VP, Barcellos C. Spatial-temporal diffusion of dengue in the municipality of Rio de Janeiro, 2000-2013. *Cadernos de Saúde Pública*. 2017; **33**(2):1–13.

[27] Lin CH, Schiøler KL, Jepsen MR, Ho CK, Li SH, Konradsen F. Dengue outbreaks in High-Income area, Kaohsiung city, Taiwan, 2003-2009. *Emerging Infectious Diseases.* 2012; **18**(10):1603–1611.

[28] Telle O, Vaguet A, Yadav NK, Lefebvre B, Daudé E, Paul RE, Cebeillac A, Nagpal BN The Spread of Dengue in an Endemic Urban Milieu–The Case of Delhi, India. *PloS One.* 2016; **11**(1):e0146

[29] Reyes-castro PA, Harris RB, Brown HE, Christopherson GL, Ernst KC. Spatiotemporal and neighborhood characteristics of two dengue outbreaks in two arid cities of Mexico. *Acta Tropica.* 2017; **167**:174–182,

[30] Tsai CT, Sung FC, Chen PS, Lin SC. Exploring the spatial and temporal relationships between mosquito population dynamics and dengue outbreaks based on climatic factors. *Stochastic Environmental Research and Risk Assessment.* 2012; **26**(5):671–680.

[31] Yu HL, Yang SJ, Yen HJ, Christakos G.A spatio-temporal climate-based model of early dengue fever warning in southern Taiwan. *Stochastic Environmental Research and Risk Assessment.* 2011; **25**(4):485–494.

[32] Yu HL, Angulo JM, Cheng MH, Wu J, Christakos G.An online spatiotemporal prediction model for dengue fever epidemic in Kaohsiung (Taiwan). *Biometrical Journal.* 2014; **56**(3):428–440.

[33] Yu HL, Lee CH, Chien LC A spatiotemporal dengue fever early warning model accounting for nonlinear associations with hydrological factors: a Bayesian maximum entropy approach. *Stochastic Environmental Research and Risk Assessment.* 2016; **30**(8):2127–2141.

[34] Lawson A, Browne WJ and Vidal Rodeiro C. *Disease Mapping with WinBUGS and MlwiN.* 2003. John Wiley & Sons Ltd. Chichester, England.

[35] Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CAS, Sá Carvalho M, Barcellos C. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers and Geosciences.* 2011; **37**(3):371–381.

[36] Huang X, Yakob L, Devine G, Frentiu FD, Fu SY, Hu W (2016) Dynamic spatiotemporal trends of imported dengue fever in Australia. *Scientific Reports.* 2016; **6**:30,360.

[37] Cadavid Restrepo A, Baker P and Clements AC.National spatial and temporal patterns of notified dengue cases, Colombia 2007–2010. *Tropical Medicine and International Health.* 2014; **19**(7): 863–871.

[38] Wijayanti SPM, Porphyre T, Chase-Topping M, Rainey SM, McFarlane M, Schnettler E, Biek R, Kohl A.The Importance of Socio-Economic Versus Environmental Risk Factors for Reported Dengue Cases in Java, Indonesia. *PLoS Neglected Tropical Diseases.* 2016; **10**(9):1–15,

[39] Lowe R, Bailey TC, Stephenson DB, Jupp TE, Graham RJ, Barcellos C and Carvalho MS. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. *Statistics in Medicine.* 2013; **32**:864–883,

[40] Lowe R, Barcellos C, Coelho CAS, Bailey TC, Coelho GE, Graham R, Jupp T, Massa Ramalho W, Stephenson DB, Rodó X. Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts. *The Lancet Infectious Diseases*. 2014; **14**(7): 619–626,

[41] Stewart-Ibarra AM, Lowe R. Climate and Non-Climate Drivers of Dengue Epidemics in Southern Coastal Ecuador. *American Journal of Tropical Medicine and Hygiene.* 2013; **88**(5):971–981,

[42] Lowe R, Cazelles B, Paul R, Rodó X. Quantifying the added value of climate information in a spatio-temporal dengue model.*Stochastic Environmental Research and Risk Assessment.* 2016; **30**(8):2067–2078,

[43] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine.* 2000; **19**:2555-2567.

[44] Villar LA, Rojas DP, Besada-Lombana S, Sarti E. Epidemiological Trends of Dengue Disease in Colombia (2000-2011): A Systematic Review. *PLoS Neglected Tropical Diseases.* 2015; **9**(3):1–16

[45] Villar L, Dayan GH , Arredondo-García JL , Rivera DM , *et al.* Efficacy of a Tetravalent dengue Vaccine in Children in Latin America. *The New England Journal of Medicine.* 2015; **372** (2):1132-123.

[46] Villabona-Arenas CJ, Ocazionez Jimenez RE, Jimenez Silva CL. Dengue Vaccine: Considerations before Rollout in Colombia. *PLoS Neglected Tropical Diseases.* 2016; **10**(6): e0004653.

[47] Departamento Administrativo Nacional de Estadśtica (DANE), Dirección de Geoestadística (National Administrative Department of Statistics. Geostatistics office) (2015) Guía de descarga y uso, marco geoestadístico nacional (Download and use guide, national geostatistical framework). `https://geoportal.dane.gov.co/metadatos/descarga_mgn/descargas/Manual_Descarga_MGN.pdf`

[48] Wan Z, Hook S, Hulley G (2015) *MOD11A2 MODIS/Terra Land Surface Temperature/Emissivity 8-Day L3 Global 1km SIN Grid V006.* NASA EOSDIS Land Processes DAAC. doi: 10.5067/MODIS/MOD11A2.006

[49] Didan, K (2015) *MOD13Q1 MODIS/Terra Vegetation Indices 16-Day L3 Global 250m SIN Grid V006.* NASA EOSDIS Land Processes doi: 10.5067/MODIS/MOD13Q1.006

[50] USGS (2011) Modis Reprojection Tool User's Manual. Release 4.1 April 2011. Land Processes DAAC. United States Geological Service (USGS) Earth Resources Observation and Science (EROS) Center.

[51] Besag J and Kooperberg C. On conditional and intrinsic autoregressions. *Biometrika*. 1995; **4**:733-46.

[52] Leroux B, Lei X, and Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In M. Halloran and D. Berry (eds), *Statistical Models in Epidemiology, the Environment and Clinical Trials*, 1999, pp. 135–78. Springer-Verlag, New York, NY.

[53] Lagazzio C, Dreassi E, Biggeri A. Hierarchical Bayesian model for space–time variation of disease risk. *Statistical Modeling.* 2001 **1**: 17–29.

[54] Lee D and Shaddick G. Time-Varying Coefficient Models for the Analysis of Air Pollution and Health Outcome Data. *Biometrics.* 1999; **63**: 1253–1261.

[55] Martínez-Bello DA, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling of time series of dengue disease case counts.. *PLoS Neglected Tropical Diseases.* 2017; **11**(7):e0005696.

[56] Lunn DJ, Thomas A, Best N, and Spiegelhalter D. WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Stat Comput.* 2000; **10**: 325-337.

[57] Lunn D, Spiegelhalter D, Thomas A, and Best N. The BUGS project: Evolution, critique, and future directions. *Statistics in Medicine.* 2009; **28**: 3049-3067.

[58] Gelman, A and Rubin, DB Inference from iterative simulation using multiple sequences, *Statistical Science*. 1992; **7**, 457-511.

[59] Watanabe S. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research.* 2010; **11**: 3571–3594.

[60] Vehtari, A., Gelman, A., Gabry, J. (2016) Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistical Computing.* doi: :10.1007/s11222-016-9696-4

[61] Vehtari A, Gelman A and Gabry J (2016) loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models R package version 0.1.6. `https://github.com/jgabry/loo`

[62] Naish S, Dale P, Mackenzie JS, McBride J, Mengersen K, Tong S. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC Infectious Diseases.* 2014; **14**(1):167.

[63] Nazri CD, Rodziah I and Hashim A. Distribution pattern of a dengue fever outbreak using GIS. *Journal of Environmental Health Research.* 2009; **9**(2): 89-97.

[64] Troyo A, Fuller DO, Calderón-Arguedas O, Solano ME, and Beier JC. Urban structure and dengue fever in Puntarenas, Costa Rica. *Singapore Journal of Tropical Geography.* 2009; **30**(2): 265–282.

[65] Buczak AL, Koshute PT, Babin SM, Feighner BH and Lewis SH.A data-driven epidemiological prediction method for dengue outbreaks using local and remote sensing data. *BMC Medical Informatics and Decision Making.* 2012; **12**:124.

[66] Araujo RV, Albertini MR, Costa-da-Silva AL, Suesdek L, Soares Franceschi NC, Bastos NM, Katz G, Cardoso VA, Castro BC, Capurro ML, Cardoso VLA.São Paulo urban heat islands have a higher incidence of dengue than other urban areas. *Brazilian Journal of Infectious Diseases.* 2015; **19**(2):146–155.

[67] Poveda G, Estrada-Restrepo OA, Morales JE, Hernández O, Galeano A and Osorio S. Integrating knowledge and management regarding the climate malaria linkages in Colombia. *Current Opinion in Environmental Sustainability.* 2011; **3**:448–460.

[68] Quintero-Herrera LL, Ramírez-Jaramillo V, Bernal-Gutiérrez S, Cárdenas-Giraldo EV *et al.* Potential impact of climatic variability on the epidemiology of dengue in Risaralda, Colombia, 2010—2011. *Journal of Infection and Public Health.* 2015; **8**: 291-297.

[69] Herbreteau V, Salem G, Souris M, Hugot JP, Gonzalez JP. Thirty years of use and improvement of remote sensing, applied to epidemiology: From early promises to lasting frustration. *Health & Place.* 2007; **13**: 400–403.

[70] Kazansky Y, Wood D, Sutherlun J. The current and potential role of satellite remote sensing in the campaign against malaria. *Acta Astronautica.* 2016; **121**:292–305.

[71] Romero-Vega L, Pacheco O, de la Hoz-Restrepo F, Díaz-Quijano FA. Evaluation of dengue fever reports during an epidemic, Colombia. *Revista de Saude Pública.* 2014; **48**(6):899–905.

[72] Ugarte MD, Adin A and Goicoa T. Two-level spatially structured models in spatio-temporal disease mapping. *Statistical Methods in Medical Research.* 2016; **25**(4) 1080–1100.

Chapter 5

Two-level resolution of relative risk of dengue disease in a hyperendemic city of Colombia

Abstract

Risk maps of dengue disease offer to the public health officers a tool to model disease risk in space and time. We analyzed the geographical distribution of relative incidence risk of dengue disease in a high incidence city from Colombia, and its evolution in time during the period January 2009 - December 2015, identifying regional effects at different levels of spatial aggregations. Cases of dengue disease were geocoded and spatially allocated to census sectors, and temporally aggregated by epidemiological periods. The census sectors are nested in administrative divisions defined as communes, configuring two levels of spatial aggregation for the dengue cases. Spatio-temporal models including census sector and commune-level spatially structured random effects were fitted to estimate dengue incidence relative risks using the integrated nested Laplace approximation (INLA) technique. The final selected model included two-level spatial random effects, a global structured temporal random effect, and a census sector-level interaction term. Risk maps by epidemiological period and risk profiles by census sector were generated from the modeling process, showing the transmission dynamics of the disease. All the census sectors in the city displayed high risk at some epidemiological period in the outbreak periods. Relative risk estimation of dengue disease using the INLA offered a quick and powerful method for parameter estimation and inference.

5.1 Introduction

Dengue is an arboviral disease caused by a *Flavivirus* belonging to the family *Flaviviridae*, which includes virus transmitted by mosquitoes such as the yellow fever virus, the Zika virus, the West Nile virus, among others. Dengue virus presents four distinct serotypes (DEN-1, DEN-2, DEN-3 and DEN-4) [1] [2], affecting people in tropical and subtropical countries in urban poor areas, suburbs and crowded neighborhoods (World Health Organization [3]). Since 2005, in the world, dengue deaths increased by 48.7% (15.1–90.9), resulting in 18400 deaths (11800–22700) in 2015 [4].

In Latin America, the increasing transmission intensity contributes to growing concerns about other viruses transmitted by *Aedes* mosquitoes, including the Chikungunya and Zika viruses [4], and emergent arboviral diseases such as Mayaro and Oropouche [5]. Racloz *et al.* [6] describe and analyze the epidemiological models attempting to predict dengue outbreaks, concluding that previous studies and modeling efforts have not sufficiently accounted for the spatio-temporal features of dengue disease in the modeling process. Louis *et al.* [7] review tools for surveillance, prevention and control of dengue focused on mapping dengue risk, finding a high diversity of dengue risk maps representing mainly descriptive and retrospective data. Naish *et al.* [8] review the spatial and spatio-temporal association of dengue disease and environmental, socioeconomic, and climatic factors. They found a diverse frame of statistical methods not integrated to useful public health systems, suggesting the need of combining research efforts to be efficient in dengue surveillance and control.

An specialized branch of disease mapping methods centers in the relative risk estimation on areal data. Relative risk corresponds to the excess (or lack) of disease risk in an area given a local and a global basal risk [9]. Relative risk could be estimated by descriptive or model based approaches. While the first option brings relatively easy and quick results, the great variability inherent to classical risk estimation measures makes necessary to use models to smooth risks using information of spatial and temporal neighbors. Model-based relative risk estimation has been carried out within a hierarchical Bayesian framework in spatial and spatio-temporal disease mapping, with generalized linear mixed models playing a major role. Knorr-Held [10] presented a framework for the spatio-temporal modeling of risks for areal data, extending the spatial model of Besag *et al.* [11].

Relative risk estimation of dengue disease has been developed using spatial and spatio-temporal data at several spatial resolutions. For example, spatial modeling of dengue data has been applied to data from Brazil [12] and Colombia [13], while spatio-temporal dengue data have been analyzed using relative risk models in Brazil [14, 20–22], Ecuador [18], Thailand [17], Colombia [19] [20] and Indonesia [21]. However, most of these analyses did not fully explore the space-time interaction effect model framework. Additive models are considered by Lowe *et al.* [14, 20–22] and Stewart-Ibarra *et al.* [18], while only the spatio-temporal unstructured interaction effect model is considered by Wijayanti *et al.* [21]. As far as we know, only Martínez-Bello *et al.* [20] consider models

of dengue relative risks with full space-time interaction terms.

Colombia is an endemic country for dengue disease. This can be attributed to high density of people living in cities and towns, with environmental and climatic conditions favouring the dengue vector development [22]. The city of Bucaramanga is one of the Colombian cities with the highest annual incidence of dengue disease in the period 2009-2015, leading the city to be one of the cities where dengue vaccines have been tested [23] [24], with some criticism from some Colombian researchers [25]. While Colombia presented an incidence rate of 624 cases per 100,000 inhabitants in 2010, Bucaramanga reported an incidence rate of 1322.1 per 100,000 inhabitants. Data on incident cases of dengue disease (dengue and severe dengue) were obtained from the SIVIGILA (Colombian public health surveillance system) for the urban area of Bucaramanga for the period from January 2009 to December 2015, and a battery of spatio-temporal models including the two-level spatially structured model proposed by Ugarte *et al.* [26] were considered.

The aim of the study is to analyze the geographical distribution of relative incidence risk of dengue disease in the city of Bucaramanga and its evolution in time during the period January 2009 - December 2015, identifying regional effects at different levels of spatial aggregations.

5.2 Materials and Methods

5.2.1 Cases of dengue disease from Bucaramanga, Colombia

Dengue cases from the city of Bucaramanga were geocoded and allocated to one of 94 census sectors. A census sector is a cartographic unit obtained from the aggregation of census sections which at the same time are the aggregation of census blocks. A census block is "*a lot of built or unbuilt land bounded by vehicular or pedestrian traffic roads of a public nature*", a census section is "*a cartographic bounded urban division approximately equal to 20 contiguous census blocks and belonging to a urban sector*", while a census sector is a "*census cartographic division at urban level, generally equivalent to a neighbor (in the principal cities), comprising between 1 to 9 census sections*" (definitions adapted from [27]). The census sectors are nested in communes. The commune is an administrative division in the municipality, representing census sectors sharing similar geographical and physical characteristics. The city of Bucaramanga covers an urban area of 27 km^2, with a population of 527,913 people (projection 2016) living in 94 census sectors nested in 17 communes. The city is located at 959 m above sea level with the coordinates $7°7'07''$ N - $73°06'58''$ W.

The inclusion of the commune as a second level of aggregation is justified by two reasons. First, a great burden of the analysis and intervention of the notification diseases at municipal level is undertaken by the health authorities employing the geographical division comprised by the communes. However, the key health event corresponding

to the dengue case occurs at house or household level. For the cases at hand, working at house or household level is challenging from the computational side, then the cases were aggregated at census sector for the sake to represent the dengue risk at small spatial scale. Secondly, as expressed above, the commune corresponds to a physical division of the city by neighborhoods and city blocks delimited by clear spatial divisions. If we include the vector biology of the disease (the mosquito *Aedes aegypti*) within the risk estimation of dengue, we could think that the vector is confined to small areas sharing special conditions for the mosquito development, which is accounted with a second level of aggregation such as the commune.

The geocoding process followed the next protocol: dengue cases data were obtained from the surveillance system of public health (SIVIGILA) for the period January 2009 to December 2015. The SIVIGILA database is an online system allowing the Colombian health institutions to register the diseases of obligatory notification. The dengue data included address, sex, age, and an identification code that anonymizes the name and personal identity of the case to the geocoder. The geocoding process started with a database checked for duplicates of 39,775 records corresponding to the notified dengue cases from health institutions in Bucaramanga. Only the records with address of residence belonging to Bucaramanga were considered, discarding cases without address, with rural address or wrong addresses. An `R` software [28] script sent batches of addresses to the web geocoding service of ArcGIS server. The web server returned JSON files, which were checked and accepted, or revised for a new geocoding cycle. At the end of the process, we succesfully geocoded to the urban area of Bucaramanga a total of 25,365 cases. Then, the coordinates obtained from the geocoding process belonging to every dengue case were allocated to census sectors using the cartography generated by the National Geostatistical Framework, 2005 [27]. In addition, the cases were temporally aggregated in epidemiological periods, composed by four epidemiological weeks, for the entire study period. The epidemiological period is the common time measure employed by the health offices in South and Central America, with a total of 91 epidemiological periods between January 2009 and December 2015 (13 epidemiological periods by year, and 7 epidemiological years).

We obtained disaggregated data by census sector, sex, and five-years age groups from the Colombian Census 2005, and calculated a cumulative crude incidence rate according to these variables. We computed cumulative expected dengue cases per area (census sectors and communes) and the seven-years period as the product of the cumulative crude incidence rate and the population at risk by age-groups and sex in every census sector and commune. Then, we added the cumulative expected cases per census sector and commune by age-group and sex, obtaining the cumulative expected cases per area. Finally, the cumulative expected dengue cases were divided by the number of epidemiological periods to obtain expected cases per area and epidemiological period.

5.2.2 Two-level spatially structured models in space-time disease mapping

Let us assume that the city of Bucaramanga is divided into n census sectors labeled as $i = 1,\ldots,n$, that are nested into m communes labeled as $j = 1,\ldots,m$. For each census sector i, data are available for different epidemiological periods labeled by $t = 1,\ldots,T$. Let O_{it}, e_{it} and r_{it} denote the number of observed dengue cases, the number of expected dengue cases and the relative risk of dengue disease for census sector i and epidemiological period t, respectively. Then, conditional on the relative risk, the number of counts is assumed to be Poisson distributed with mean $\mu_{it} = e_{it} r_{it}$, that is,

$$O_{it}|r_{it} \sim \text{Poisson}(\mu_{it} = e_{it} r_{it})$$

$$\log \mu_{it} = \log e_{it} + \log r_{it}.$$

Depending on the specification of $\log r_{it}$ several models could be defined. Most of the research in space-time disease mapping is based on conditional autoregressive (CAR) priors for both spatial and temporal effects (Knorr-Held [10]). Extensions of these models were proposed by Ugarte *et al.* [26] for analyzing small area data that are naturally grouped into larger regions. The models include two-level of spatially structured random effects, identifying regional effects and modeling space-time interactions at different levels of spatial aggregations. In what follow, we briefly describe some of these models.

First, a model with census sector level space-time interaction has been considered (hereafter *TL-Model A*), where the log-risk is modeled as

$$\log r_{it} = \eta + \xi_i + \psi_{j(i)} + \gamma_t + \delta_{it}, \tag{5.1}$$

where $j(i)$ denotes that census sector i belongs to the commune $j = 1,\ldots,m$. Here η is an intercept representing an overall level of risk, ξ_i and $\psi_{j(i)}$ are census sector and commune level spatially structured random effects respectively, γ_t is a temporally structured random effect, and δ_{it} is the space-time interaction effect that models the dependence between the census sectors and the epidemiological periods. If the interaction term is dropped, an additive model is obtained. A Leroux *et al.* [29] CAR (LCAR) prior distribution is given to both spatial random effects, that is,

$$\xi = (\xi_1,\ldots,\xi_n)' \sim N(\mathbf{0}, [\tau_\xi(\lambda_\xi \mathbf{R}_\xi + (1-\lambda_\xi)\mathbf{I}_n)]^{-1}),$$

$$\psi = (\psi_1,\ldots,\psi_m)' \sim N(\mathbf{0}, [\tau_\psi(\lambda_\psi \mathbf{R}_\psi + (1-\lambda_\psi)\mathbf{I}_m)]^{-1}),$$

where τ_ξ and τ_ψ are precision parameters, λ_ξ and λ_ψ are spatial smoothing parameters taking values between 0 and 1, $\mathbf{I}_n$ and $\mathbf{I}_m$ are identity matrices of dimension $n \times n$ and $m \times m$ respectively, $\mathbf{R}_\xi$ is the $n \times n$ neighborhood matrix of the census sectors, and $\mathbf{R}_\psi$ is the $m \times m$ neighborhood matrix of the communes. Note that spatial independence is assumed when the spatial smoothing parameters are equal to zero, while intrinsic CAR prior distributions are considered when these parameters are equal to one. A first order random walk (RW1) prior distribution is given for the temporally structured random effect, that is,

$$\gamma = (\gamma_1,\ldots,\gamma_T)' \sim N(\mathbf{0}, [\tau_\gamma \mathbf{R}_\gamma]^-).$$

Here τ_γ is a precision parameter and $\mathbf{R}_\delta$ is the $T \times T$ structure matrix of a RW1.

Finally, the following prior distribution is assumed for the space-time interaction random effect $\delta = (\delta_{11}, \ldots, \delta_{1T}, \ldots, \delta_{n1}, \ldots, \delta_{nT})'$

$$\delta \sim N(\mathbf{0}, [\tau_\delta \mathbf{R}_\delta]^-).$$

Here τ_δ is a precision parameter and $\mathbf{R}_\delta$ is the $nT \times nT$ matrix obtained as the Kronecker product of the corresponding spatial and temporal structure matrices. Note that a commune level interaction effect can be also considered in the model of Equation (5.1), modeling the log-risks as (hereafter *TL-Model B*)

$$\log r_{it} = \eta + \xi_i + \psi_{j(i)} + \gamma_t + \delta_{j(i)t}. \tag{5.2}$$

As proposed by Knorr-Held [10], four types of space-time interactions can be defined for TL-Model A and TL-Model B (see Table 5.1).

A sensible modification of these models is to account for spatial variability only among those census sectors belonging to the same commune. In this case, the census sector level random effects are distributed as $\xi^* \sim N(\mathbf{0}, [\tau_\xi(\lambda_\xi \mathbf{R}_\xi^* + (1-\lambda_\xi)\mathbf{I}_n)]^{-1})$, where $\mathbf{R}_\xi^* = \text{blockdiag}(\mathbf{R}_{\xi_1}, \ldots, \mathbf{R}_{\xi_m})$ is a block-diagonal matrix and $\mathbf{R}_{\xi_j}$ is the neighborhood matrix of census sectors within the jth commune. Both census sector or commune level space-time interactions can be considered, defining the following models

$$\begin{aligned} &\textit{TL-Model C:} \qquad \log r_{it} = \eta + \xi_i^* + \psi_{j(i)} + \gamma_t + \delta_{it}^*, \\ &\textit{TL-Model D:} \qquad \log r_{it} = \eta + \xi_i^* + \psi_{j(i)} + \gamma_t + \delta_{j(i)t}. \end{aligned} \tag{5.3}$$

Again, four different types of space-time interaction can be defined for the models of Equation (5.3), obtained as the Kronecker product of the corresponding spatial and temporal structure matrices.

5.2.3 Model inference and estimation

Different spatio-temporal models of relative risk described above were fitted using the integrated nested Laplace approximation (INLA) technique, an approximate method for Bayesian inference for latent Gaussian models developed by Rue *et al.* [30]. INLA provides reliable results in short computational time when the precision matrices of the random effects are sparse, allowing to make Bayesian inference without running long and complex Markov chain Monte Carlo (MCMC) algorithms. This technique can be used in the free statistical software `R` through the `R-INLA` package. Appropriate identifiability constraints have been considered for each model, which are derived by reparameterizing the random effects using the spectral decomposition of their precision matrices (see Goicoa *et al.* [31]). Non-informative prior distributions were assigned to the model

Table 5.1: Specification for the different types of space-time interactions.

Interaction		Structure	
Type	$\mathbf{R}_\delta$	Spatial	Temporal
	Two Level-Model A		
I	$\mathbf{I}_n \otimes \mathbf{I}_T$	–	–
II	$\mathbf{I}_n \otimes \mathbf{R}_\gamma$	–	✓
III	$\mathbf{R}_\xi \otimes \mathbf{I}_T$	✓	–
IV	$\mathbf{R}_\xi \otimes \mathbf{R}_\gamma$	✓	✓
	Two Level-Model B		
I	$\mathbf{I}_m \otimes \mathbf{I}_T$	–	–
II	$\mathbf{I}_m \otimes \mathbf{R}_\gamma$	–	✓
III	$\mathbf{R}_\psi \otimes \mathbf{I}_T$	✓	–
IV	$\mathbf{R}_\psi \otimes \mathbf{R}_\gamma$	✓	✓

hyperparameters as follows

$$\begin{aligned} \eta &\sim \text{Normal}(0, 1000), \\ \lambda_\xi, \lambda_\psi &\sim \text{Uniform}(0, 1), \\ \frac{1}{\sqrt{\tau_\xi}}, \frac{1}{\sqrt{\tau_\psi}}, \frac{1}{\sqrt{\tau_\gamma}}, \frac{1}{\sqrt{\tau_\delta}} &\sim \text{Uniform}(0, \infty). \end{aligned}$$

Some model selection criteria were considered to compare the different models in terms of model fitting and complexity. The deviance information criterion (DIC) (Spiegelhalter *et al.* [32]) is the most commonly used measure of model fit based on the deviance for Bayesian models, which is computed as the sum of the posterior mean of the deviance $\bar{D}$ (a measure of goodness of fit) and the number of effective parameters p_D (a measure of model complexity). Although the use of the DIC has been widespread during the last years, it has been criticized by several authors in the literature. It is recognized that the DIC values may underpenalize complex models containing random effects in disease mapping, so the corrected version of the DIC proposed by Plummer [33] was also considered in this paper. It is also known that the DIC can produce negative estimates of the effective number of parameters in a model. Some authors recommend the use of the Watanabe-Akaike information criterion (WAIC) (Watanabe [34]) instead of the DIC (see for example, Gelman *et al.* [35]; Vehtari *et al.* [36]). The WAIC criterion was also computed here. Finally, we provide the cross-validate logarithmic score (LS) (Gneiting and Raftery [37]) as a criterion based on the model posterior predictive distribution.

5.3 Results

5.3.1 Summary statistics

A total of 25,365 dengue cases were successfully geocoded to the city area of Bucaramanga. As shown in Figure 5.1A, three main outbreaks were experienced in the city during the period January 2009 to December 2015: in the first semester of 2010 with around 940 cases, and in the first semester of 2013 and 2014 presenting near to 550 cases each. Figure 5.1B shows the age-groups [5-9] and [10-14] years presenting the highest annual average cumulative incidence of dengue disease for the study period (1,349 and 1,238 cases by 100,000 inhabitants, respectively). The maximum number of dengue cases per census sector and commune were 47 and 97 cases respectively.

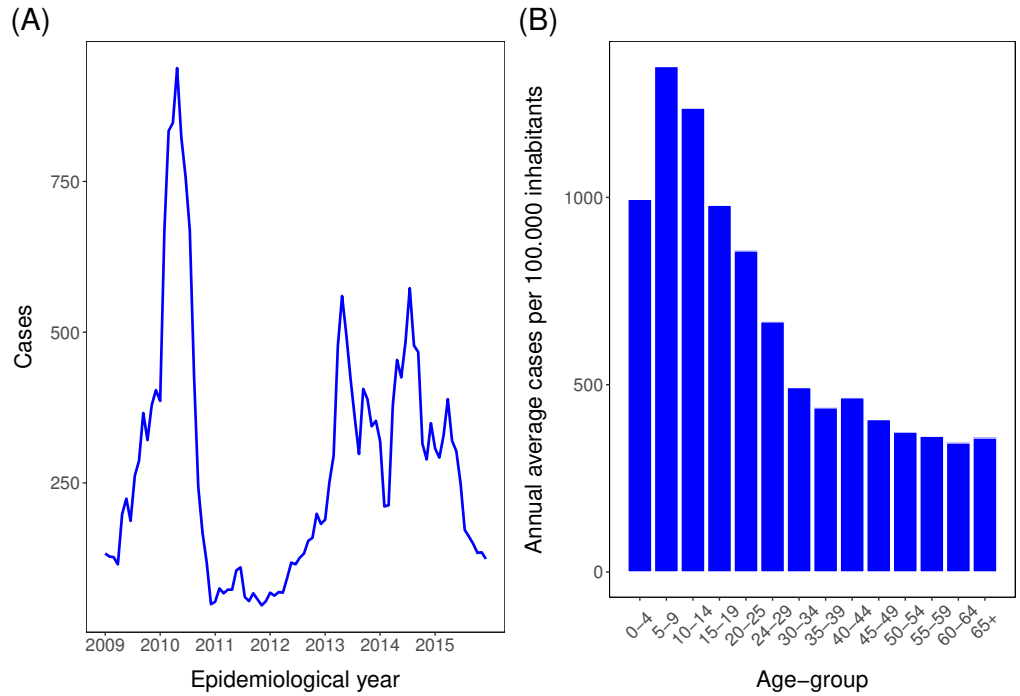


Figure 5.1: Descriptive analysis of dengue disease cases in the city of Bucaramanga, Colombia. (A) Cases by epidemiological period. (B) Annual average cases per 100.000 inhabitants and age-groups.

Figure 5.2 provides the cumulative standardized incidence rate (SIR) of dengue disease in the 293 census sectors and 94 communes for the 7-year time period (2009-2015). The cumulative SIR of dengue is an indirect method of adjustment for age and sex, acting as a measure to compare dengue cases in each area and time point with the whole city during the study period. The cumulative SIR per census sector (Figure 5.2A) shows a diffuse incidence pattern with a few high incidence census sectors to the west of the city, while the cumulative SIR per commune (Figure 5.2B) reveals high incidence to

the south and central communes of the city.

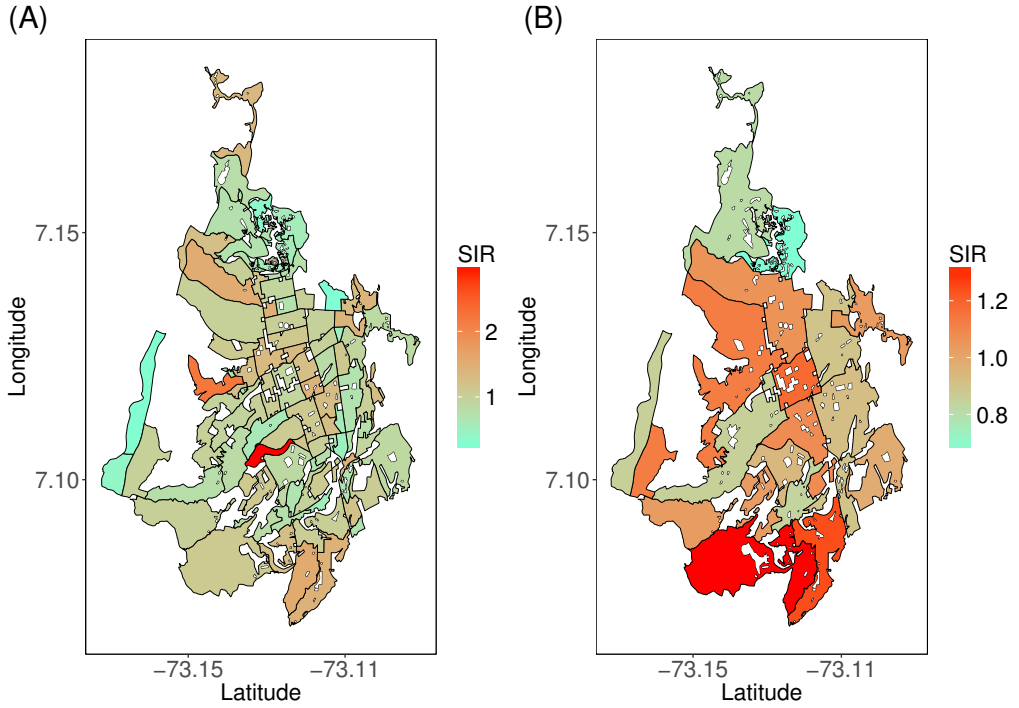


Figure 5.2: Cumulative standardized incidence rates (SIRs) of dengue disease by communes and census sectors. (A) SIR of dengue cases by census sector. (B) SIR of dengue cases by commune.

5.3.2 Results from the selected model

Table 5.2 shows the results from fitting the models described above with `R-INLA` using the simplified Laplace approximation strategy. In general, the models with census sector level interaction effect are better than those considering a commune level interaction effect. Nevertheless, from a computational point of view, the latter models are much faster because the space-time precision matrix ($\mathbf{R}_\delta$) has lower dimension and less identifiability constraints are needed. In addition, according to the different model selection criteria, the model with the usual spatial neighborhood structure performs better than TL-Models C and D that incorporate a more complex neighborhood structure between areas. As reported in Table 5.2, TL-Model A with completely structured (Type IV) interaction random effect shows the lowest values for all the model selection criteria considered here (almost 80 units less than TL-Model A and C with Type II interaction effect in terms of DIC and WAIC, and 150 units less in terms of corrected DIC).

Finally, Table 5.2 shows that the number of effective parameters decreases for Type

II and IV interaction models in comparison with Type I and III. This might seem counter-intuitive since a Type IV interaction model is more complex in terms of the covariance structure induced for the space-time neighboring points. However, we note that the number of effective parameters is also an indicator of the degree of smoothness induced by the model. As the random effects of the model induce more smoothness, i.e., as the shrinkage towards zero (the mean) is stronger, the more we move away from the saturated model, and therefore the model is less complex. This seems the be the reason why models with Type I or III interaction random effects, that do not induce smoothing effects between temporal neighbors, shows higher values of p_D.

Table 5.3 shows the summary statistics for the precision parameters from the selected model. The posterior mean of the spatial smoothing parameter of the LCAR prior distribution for the census sector random effect (λ_ξ) is 0.283, which is interpreted as small spatial dependence between these areas. For the commune random effect, the posterior mean of the spatial smoothing parameter (λ_ψ) is 0.448, indicating a moderate spatial dependence between communes in the same study period.

By fitting spatio-temporal models with two-level of spatial random effects, we provide a tool to establish the association between the commune and the census sector with the relative risk of dengue disease, accounting for those geographical factors specific to the area covered by the commune. Figure 5.3 shows the maps for the census sector and commune level spatial incidence risk patterns (constant during the whole period) derived from the selected model. These spatial patterns, can be interpreted as the specific contribution of the area to the increase/decrease of the relative risks r_{it}. Figure 5.3A exposes some of the census sector located in the central areas of the city showing a large mean spatially structured pattern. The probability of the census sector spatial effects being greater than one is represented in Figure 5.3B. At commune level, a large spatial incidence patterns is observed in the southern and western communes of the city (Figure 5.3C), which is better inferred by the posterior exceedence probabilities $\mathrm{P}(\exp(\psi_{j(i)}) > 1|\mathbf{O})$ represented in Figure 5.3D.

Including two-level random effects in the model allow us to identify those census sectors and/or communes that have a significant effect on the relative risk. For example, the commune located further north in the city of Bucaramanga it is not a high risk area, but some of its census sectors show a high probability that the spatial effect is significantly higher in comparison with the whole of the sectors (see Figure 5.3). In this way, we are able to identify those high/low risk areas that show behaviors associated to both levels of spatial aggregation.

The posterior mean temporal trend and 95% credible intervals by epidemiological period (common to all areas) is shown in Figure 5.4, recovering the high risk pattern of dengue disease in the first semesters of 2010, 2013, 2014, and 2015.

Mapping the relative risk estimates of dengue disease is one of the main outputs from the modeling process. We have chosen the epidemiological periods 1 to 8 from the year 2013 to display the estimated posterior mean values of the relative risk of dengue disease (Figure 5.5). Using the relative risk implies that the one is the basal risk. The maps in

Table 5.2: Model selection criteria for the best fitted models in INLA: mean deviance ($\bar{D}$), number of effective parameters (p_D), deviance information criterion (DIC), corrected DIC (DICc), Watanabe-Akaike information criterion (WAIC) and logarithmic score (LS).

TL-Model A: $\log r_{it} = \eta + \xi_i + \psi_{j(i)} + \gamma_t + \delta_{it}$

Space-time interaction	$\bar{D}$	p_D	DIC	DICc	WAIC	LS
Additive model	30256.0	162.8	30418.8	30423.1	30552.6	15276.8
Type I	26300.3	2282.4	28582.7	30711.9	28604.0	14816.0
Type II	26792.8	1266.2	28059.0	28406.8	28251.7	14196.6
Type III	26762.0	1733.1	28495.1	29652.5	28709.7	14706.5
Type IV	26885.7	1095.4	**27981.1**	**28256.1**	**28167.3**	**14138.8**

TL-Model B: $\log r_{it} = \eta + \xi_i + \psi_{j(i)} + \gamma_t + \delta_{j(i)t}$

Space-time interaction	$\bar{D}$	p_D	DIC	DICc	WAIC	LS
Type I	28451.1	851.1	29302.2	29521.0	29713.3	14917.3
Type II	28526.7	578.3	29105.0	29186.1	29388.4	14711.6
Type III	28560.7	821.9	29382.5	29587.5	29794.5	14955.6
Type IV	28572.1	575.1	29147.2	29231.1	29431.6	14734.1

TL-Model C: $\log r_{it} = \eta + \xi_i^* + \psi_{j(i)} + \gamma_t + \delta_{it}^*$

Space-time interaction	$\bar{D}$	p_D	DIC	DICc	WAIC	LS
Additive model	30256.7	162.9	30419.6	30424.0	30553.3	15277.1
Type I	26307.5	2276.7	28584.2	30699.1	28608.9	14814.8
Type II	26789.2	1269.2	28058.3	28407.9	28249.9	14196.1
Type III	28210.7	1467.5	29678.3	30453.5	30263.7	15844.3
Type IV	28255.5	899.1	29154.6	29328.6	29514.9	14790.5

TL-Model D: $\log r_{it} = \eta + \xi_i^* + \psi_{j(i)} + \gamma_t + \delta_{j(i)t}$

Space-time interaction	$\bar{D}$	p_D	DIC	DICc	WAIC	LS
Type I	28451.3	851.8	29303.1	29522.0	29714.2	14917.8
Type II	28527.7	578.4	29106.1	29187.2	29389.5	14712.2
Type III	28561.9	821.6	29383.5	29588.1	29795.3	14955.9
Type IV	28573.2	575.1	29148.3	29232.1	29432.7	14734.7

Figure 5.5 show that in 2013, the EP 1 and 2 present a low overall relative risk in most of the census sectors, but afterwards, the relative risk spread from the center of the city in EP 3 and 4 to the rest of census sectors in EP 5 and 6, and finally decreasing slightly in the EP 7 and 8. To detect the areas with high relative incidence risk, maps of the posterior excedance probabilities $P(r_{it} > 1|\mathbf{O})$ by census sector and epidemiological period have been represented in Figure 5.6. This posterior probability distribution provides a kind of Bayesian p-value, which it could be used to detect or highlight high risk areas based on

Table 5.3: Summary statistics for the precision parameters of the TL-Model A with type IV interaction effect for the relative risk of the Dengue, Jan 2009 - Dec 2015.

Parameter	Mean	SD	Q 0.025	Q 0.5	Q 0.975
τ_ξ	6.08	1.93	3.31	5.73	10.81
λ_ξ	0.28	0.156	0.058	0.27	0.60
τ_ψ	58.82	78.04	6.72	35.59	253.68
λ_ψ	0.44	0.24	0.06	0.43	0.89
τ_γ	19.20	3.23	13.51	18.99	26.16
τ_δ	14.19	1.18	12.04	14.14	16.66

the definition of a cut point by the analyst.

Finally, we have selected eight census sector distributed across the city to plot their specific temporal evolution of dengue incidence risk during the period Jan 2009 - Dec 2015, and the posterior mean values of the estimated relative risks and 95% credible intervals by epidemiological period (Figure 5.7). Four census sectors correspond to the central areas of the city (central east area: *Cabecera* sector; central north area: *San Francisco* sector; central south area: *Real de Minas* sector; and central west area: *Campohermoso* sector), and four census sectors from the east (*Morrorico* sector), north (*Kennedy* sector), south (*Provenza* sector) and west (*Girardot* sector) areas of the city. Although the temporal evolution of relative risks are similar to the main temporal pattern represented in (Figure 5.4), subtle differences are revealed between census sectors. The main outbreak of dengue cases observed in the city of Bucaramanga (first semester of 2010) did not equally affect to all areas, observing significantly higher spikes in *Provenza*, *San Francisco*, and *Morrorrico* sectors. In addition, quite different relative risk evolutions are observed during the period 2013 to 2015. Sectors located in the central areas of the city show much more moderate risks during the last years of the analyzed period than the areas located in the suburbs of the city, where significantly high relative risks are observed in *Provenza* and *Girardot* sectors.

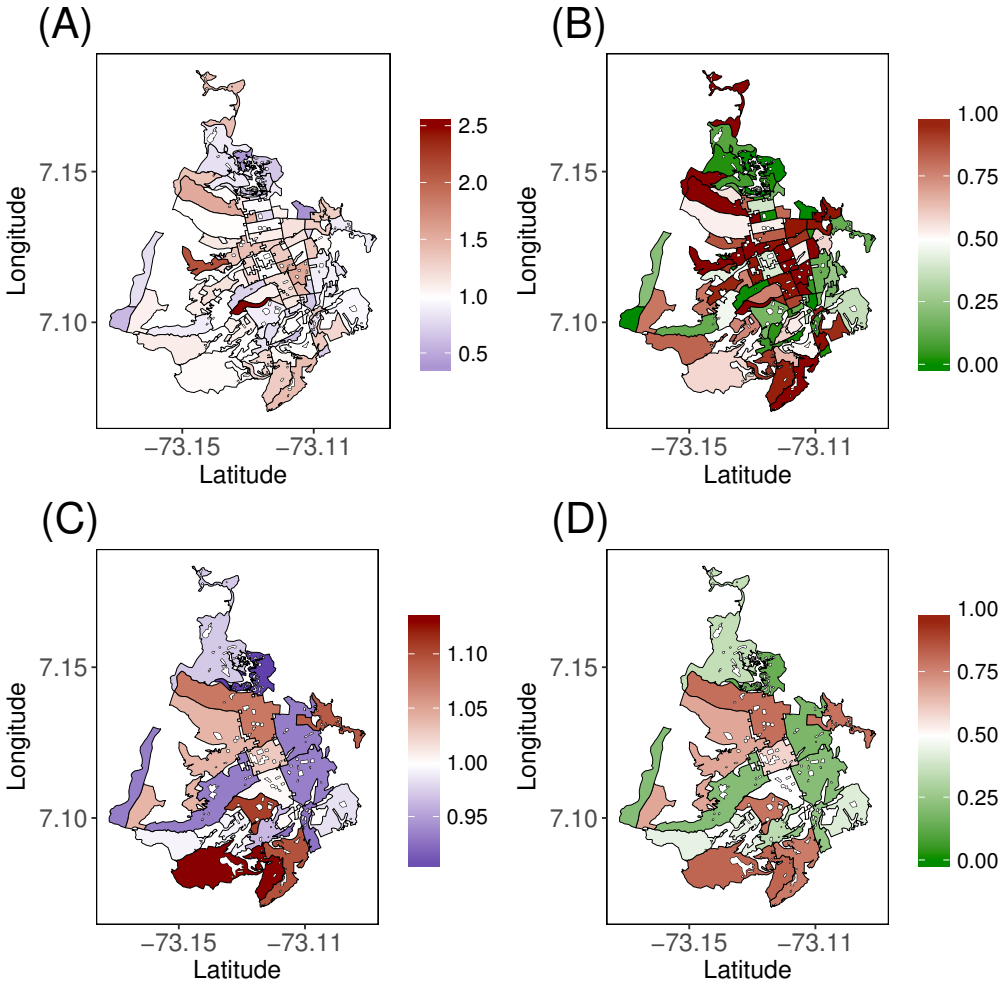


Figure 5.3: Posterior mean estimates of spatial random effects at both census sector and commune-level, and posterior excedance probability of being greater than one. (A) Map of census sector level spatial incidence risk pattern $\exp(\xi_i)$. (B) Posterior probability distribution $P(\exp(\xi_i) > 1|\mathbf{O})$. (C) Map of commune level spatial incidence risk pattern $\exp(\psi_{j(i)})$. (D) Posterior probability distribution $P(\exp(\psi_{j(i)}) > 1|\mathbf{O})$.

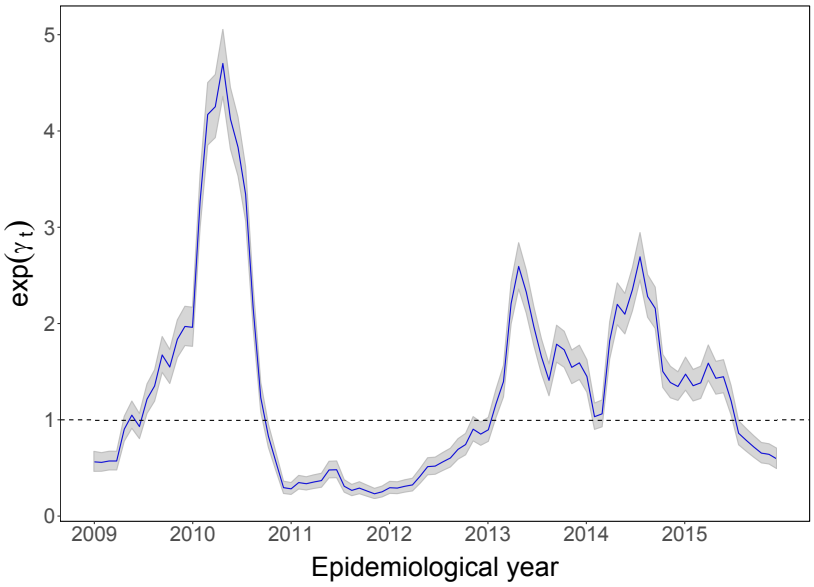


Figure 5.4: Overall temporal trend of dengue disease incidence relative risk by epidemiological period, $\exp(\gamma_t)$, and 95% credibility intervals.

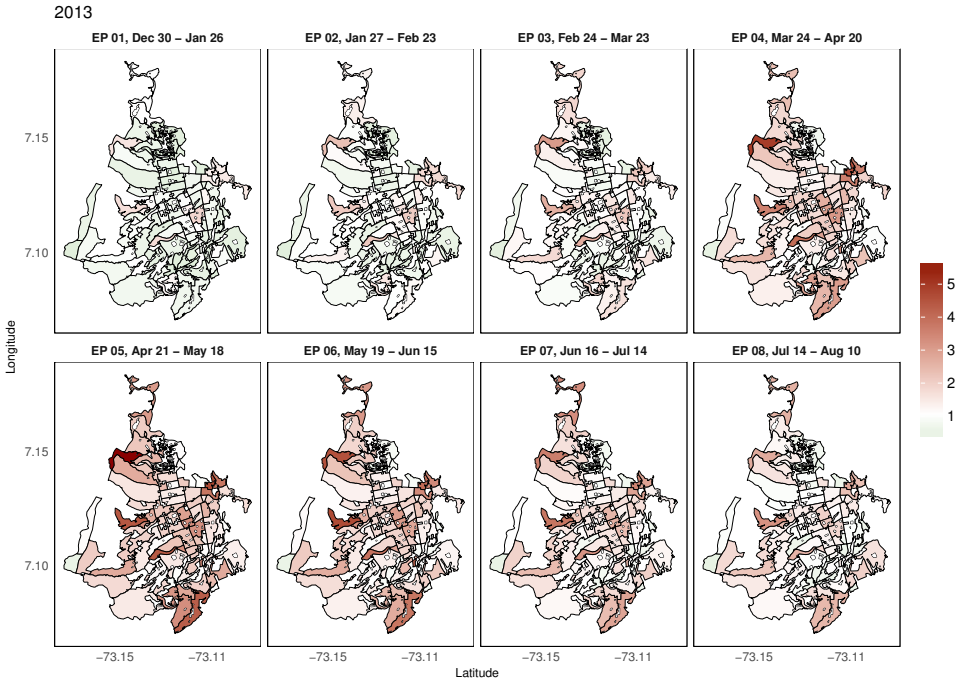


Figure 5.5: Maps with the estimated posterior mean values of the relative risk r_{it} of dengue disease by census sector for the epidemiological periods 1 to 8 of 2013.

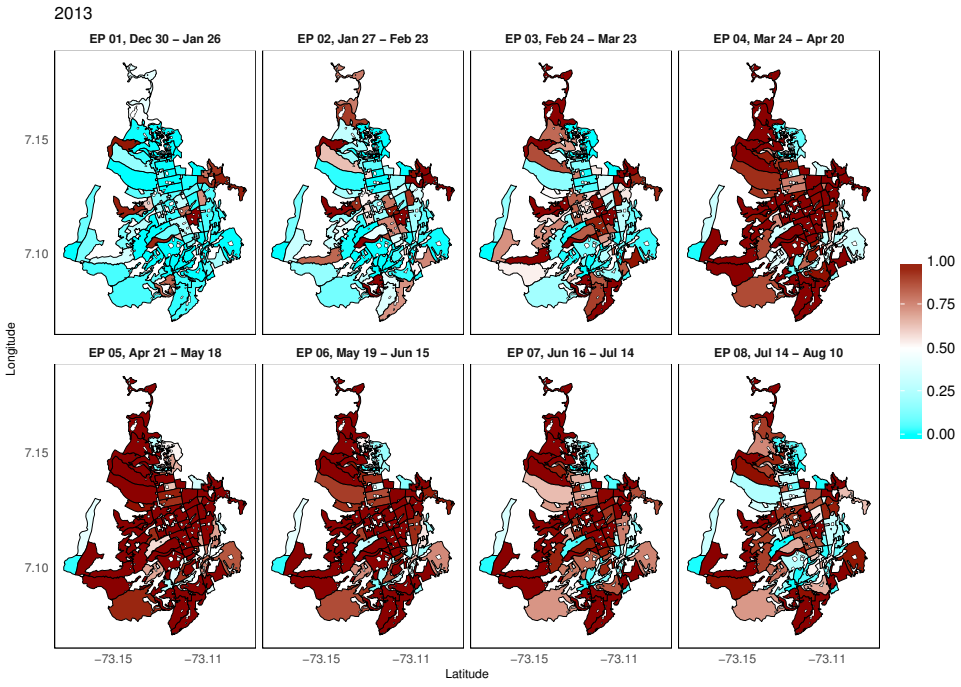


Figure 5.6: Maps of the posterior probability distribution $P(r_{it} > 1|\mathbf{O})$ of dengue disease by census sector for the epidemiological periods 1 to 8 of 2013.

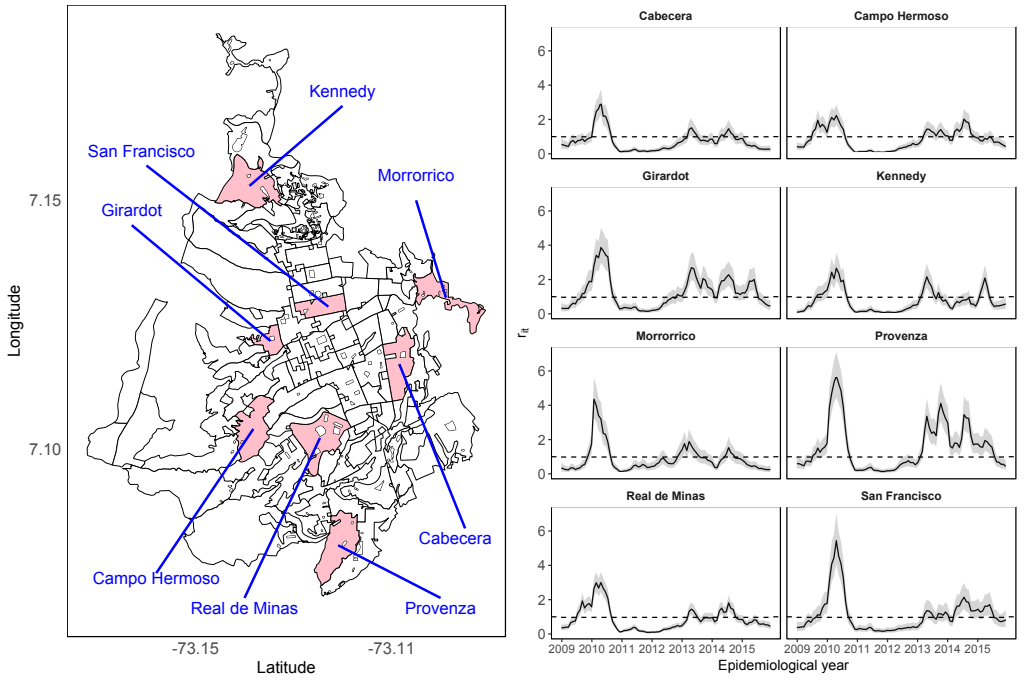


Figure 5.7: Map of selected census sector to display relative risk of dengue disease for the period Jan 2009 - Dec 2015 (left panel), and specific temporal evolution of the posterior mean estimates of relative risk and 95% credible intervals (right panel)

5.4 Discussion

In this work, spatio-temporal disease mapping models are applied to dengue incidence data in the city of Bucaramanga (Colombia). Dengue cases are spatially aggregated at two different administrative levels, the census sectors and communes, and temporally aggregated in epidemiological period. This is the first report where models with two-level of spatially structured random effects have been used to estimate dengue disease incidence relative risks in small areas. These models permit to identify regional effects at each level of spatial aggregation, considering space-time interaction effects at census sector or commune level.

We found for the particular data at hand that the best model to report the results includes spatially structured random effects for both census sector and commune levels, a temporally structured random effect for the epidemiological periods, and a completely structured interaction term over the census sectors (TL-Model A with type IV interaction). Different model selection criteria has been used to compare the behaviour of the fitted models, such as the deviance information criterion, the Watanabe-Akaike information criterion or the cross-validate logarithmic score. The selected model implies that the risk of dengue disease on every census sector is highly associated to the neighboring census sectors in space and time, where the commune effect plays an important role in the dynamics of the transmission of dengue disease across the city. The model's output of the model allows to create risk maps by epidemiological period, and risk profiles by census sector through the study period. Inspection of the risk maps permits to detect high/low risk areas in comparison with all the census sectors of the city across the period Jan 2009 - Dec 2015. Our analysis shows differences in dengue incidence risk between sectors situated in central areas of the city compared to sectors located in the suburbs of the city.

The present analysis extends the results shown by Martínez-Bello *et al.* [38], where the time series of dengue disease in the city of Bucaramanga were analyzed in a weekly basis, including covariates. The overall and sector-specific temporal trends of dengue disease obtained from the spatio-temporal model allow the comparison of the longitudinal profiles of dengue risk per sector.

The main novelty in the present work is the inclusion of the commune effect as a second level of spatial aggregation in the modeling process. The primary benefit of considering models with two-level spatial random effects is that they provide key information to the public health policy-makers, such as the geographical distribution of dengue disease relative risks by census sector, and the contribution of the communes to the decrease/increas of these risks. Models with a single level of spatial dependence were considered in Martinez *et al.* [20] to analyze dengue incidence data in the city of Bucaramanga at smaller spatial aggregation units (census sections), including covariates obtained from satellite data. However, particular attention must be paid on solving identifiability issues in disease mapping models when covariates are included, because ignoring the spatial or temporal correlation between covariates and the random effects can lead to misleading results due to confounding issues (see for example, Reich *et al.* [39];

Hodges and Reich [40]; Goicoa *et al.* [31]). It is matter of further research not only how to deal with covariates in a spatio-temporal model, but also how to do it in a model that includes spatial random effects at two-level of spatial aggregation and space-time interactions at both first or second-level area (census sectors and communes in our data of analysis).

The models are fitted using the recently derived INLA estimation technique, reducing the computational time in comparison with models fitted using MCMC simulations. The model complexity requires a great amount of time using MCMC while INLA offer a fast alternative, which it could be applied to programs of real-time spatiotemporal representation of dengue risk. In addition, the results from the models applied in the present report could also be used in combination with other statistical methods of spatial and temporal risk representation. For example, space-time clustering methods have been also applied to dengue data by Fuentes-Vallejo [41] in a hyperendemic colombian city located at the center of the country, concluding that there were not specific areas in the city driving the transmission of dengue disease, a characteristic displayed in our results.

As limitations of the study, we account that the use of notification data can lead to underrepresentation of those cases that are managed at home, without reporting to the surveillance system [42]. Also, some of the addresses possibly were not correctly geocoded, or not geocoded at all, due to mistakes in filling the notification form. Quantifying the percentage of correctly geocoded cases is difficult, although we keep out of the final data those addresses with inconsistent data. In addition, some of the cases reported as dengue were not confirmed by laboratory, but confirmed by clinical diagnosis, leading to a bias difficult to quantify in our results [42]. Furthermore, a current problem for the epidemiological studies at block, section or sector level in Colombia is the lack of updated data from the official statistics, leading to an additional source of bias on the results. Although we report these limitations, we also address that dengue is a highly recognized disease within the medical staff (physicians, nurses, public health personal) in Colombia. The epidemiological and statistical tools like the relative risk models shown in this study can help to decrease the dengue burden, by providing risk maps and risk profiles, which in first stages would be approximated to the unknown field situation, and in second stages with the addition of high quality data they will support an integrated approach to dengue surveillance and control activities with the addition of high quality data [43].

We think that further work is needed to make available to the public health policy-makers epidemiological tools to generate real-time dengue disease incidence risk maps, including environmental risk factors (rainfall, humidity, temperature, ...) or other potential explanatory variables such as vectorial ecology in the modeling process.

5.5 References

[1] Muller DA, Depelsenaire ACI, Young PR. Clinical and Laboratory Diagnosis of Dengue Virus Infection. *The Journal of Infectious Diseases* 2017; **215**: S89-S95.

[2] Ramos-Castañeda J, Barreto dos Santos F, Martínez-Vega R, Galvão de Araujo J, Joint G, Sarti E. Dengue in Latin America: Systematic Review of Molecular Epidemiological Trends. *PLoS Neglected Tropical Diseases.* 2017; **11**(1): 0005224.

[3] World Health Organization. *Global Strategy for Dengue Prevention and Control, 2010-2020.* World Health Organization, Geneva, Switzerland, 2012.

[4] Wang H, Naghavi M, Allen C, et al. Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980–2015: a systematic analysis for the Global Burden of Disease Study 2015. *The Lancet* 2016; **388**(10053):1459–1544.

[5] Alva-Urcia C, Aguilar-Luis MA, Palomares-Reyes C, Silva-Caso W, Suarez-Ognio L, Weilg P, et al. Emerging and reemerging arboviruse s: A new threat in Eastern Peru. *PLoS ONE* 2017; **12**(11): e0187897.

[6] Racloz V, Ramsey R, Tong S, Hu W. Surveillance of dengue fever virus: A review of epidemiological models and early warning systems. *PLoS Neglected Tropical Diseases* 2012; **6**(5):e1648.

[7] Louis VR, Phalkey R, Horstick O, Ratanawong P, Wilder-Smith A, Tozan Y, Dambach P. Modeling tools for dengue risk mapping - a systematic review. *International Journal of Health Geographics* 2014; **13**(1):50.

[8] Naish S, Dale P, Mackenzie JS, McBride J, Mengersen K, Tong S. Climate change and dengue: a critical and systematic review of quantitative modelling approaches. *BMC Infectious Diseases* 2014; **14**(1): 167.

[9] Lawson A. Bayesian Disease Mapping : Hierarchical Modeling in Spatial Epidemiology. Chapman & Hall/CRC interdisciplinary statistics series. Boca Raton, FL. 2009.

[10] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine* 2000; **19**:2555–2567.

[11] Besag J, York J, Mollie A. Bayesian image restoration with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* 1991; **43**(1):1–59.

[12] Ferreira G, Schmidt A. Spatial modelling of the relative risk of dengue fever in Rio de Janeiro for the epidemic period between 2001 and 2002. *Brazilian Journal of Probability and Statistics* 2006; **20**:29–47.

[13] Martínez-Bello DA, López-Quílez A, Torres Prieto A. Relative risk estimation of dengue disease at small spatial scale. *International Journal of Health Geographics* 2017; **16**:31.

[14] Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CAS, Carvalho M, Barcellos C. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers and Geosciences* 2011; **37**(3): 371–381.

[15] Lowe R, Bailey T, Stephenson D, Jupp T, Graham R, Barcellos C, Carvalho M. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. *Statistics in Medicine* 2013; **32**: 864–883.

[16] Lowe R, Barcellos C, Coelho C, Bailey T, Coelho G, Graham R, Jupp T, Massa-Ramalho W, Stephenson D, Rodó X. Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts. *The Lancet Infectious Diseases* 2014; **14**(7): 619–626.

[17] Lowe R, Cazelles B, Paul R, Rodó X. Quantifying the added value of climate information in a spatio-temporal dengue model. *Stochastic Environmental Research and Risk Assessment* 2016; **30**(8):2067–2078.

[18] Stewart-Ibarra A, Muñoz Á, Ryan S, et al. Spatiotemporal clustering, climate periodicity, and social-ecological risk factors for dengue during an outbreak in Machala, Ecuador, in 2010. *Infectious Diseases* 2014; **14**(1):610.

[19] Cadavid-Restrepo, A, Baker P, Clements ACA. National spatial and temporal patterns of notified dengue cases, Colombia 2007-2010. *Tropical Medicine & International Health* 2014; **19**(7) :863–871.

[20] Martínez-Bello D, López-Quílez A, Prieto AT. Spatiotemporal modeling of relative risk of dengue disease in Colombia. *Stochastic Environmental Research and Risk Assessment* 2017. doi: `http://doi.org/10.1007/s00477-017-1461-5`

[21] Wijayanti SPM, Porphyre T, Chase-Topping M, Rainey SM, McFarlane M, Schnettler E, Biek R, Kohl A. The Importance of Socio-Economic Versus Environmental Risk Factors for Reported Dengue Cases in Java, Indonesia. *PLoS Neglected Tropical Diseases* 2016; **10**(9): 1–15.

[22] Villar LA, Rojas DP, Besada-Lombana S, Sarti E. Epidemiological Trends of Dengue Disease in Colombia (2000-2011): A Systematic Review. *PLoS Neglected Tropical Diseases* 2015 **9**(3): e0003499.

[23] Villar L, Dayan GH, Arredondo-Garcia JL, et al. Efficacy of a tetravalent dengue vaccine in children in Latin America. *N Engl J Med* 2015; **372**:113-23.

[24] Gustavo Olivera-Botello G, Coudeville L, Fanouillere K, Guy B, Chambonneau L, Noriega F, and Jackson N. Tetravalent Dengue Vaccine Reduces Symptomatic and Asymptomatic Dengue Virus Infections in Healthy Children and Adolescents Aged 2-16 Years in Asia and Latin America. *Journal of Infectious Diseases* 2016; **214**:994-1000.

[25] Villabona-Arenas CJ, Ocazionez Jimenez RE, Jimenez Silva CL. Dengue Vaccine: Considerations before Rollout in Colombia. *PLoS Neglected Tropical Diseases* 2016 **10**(6): e0004653.

[26] Ugarte M, Adin A, Goicoa T. Two-level spatially structured models in spatio-temporal disease mapping. *Statistical Methods in Medical Research* 2016; **25**: 1080–1100.

[27] Departamento Administrativo Nacional de Estadística, Colombia. Dirección de Geoestadística. (National Administrative Department of Statistics, Colombia. Direction of Geostatistics). Capa del Nivel de Sector Urbano (urban Sector Level Layer). 2005. Marco Geoestadístico Nacional (National Geostatistical Framework). `http://www.dane.gov.co/`

[28] R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing Vienna, Austria 2016. R Foundation for Statistical Computing. `https://www.R-project.org/`

[29] Leroux BG, Lei X, Breslow N.: Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: MH, Berry.D. (eds.) Statistical Models in Epidemiology, the Environment and Clinical Trials pp.:179–191. Springer, New York 1999.

[30] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations . *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; **71**:319–392.

[31] Goicoa T, Adin A, Ugarte MD, Hodges JS. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results . *Stochastic Environmental Research and Risk Assessment.* 2018; **32**:749–770.

[32] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *Journal of Royal Statistical Society, Series B (Statistical Methodology)* 2002; **64**:583–639.

[33] Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics* 2008; **9**:523–539.

[34] Watanabe S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research* 2010; **11** :3571–3594

[35] Gelman A, Hwang J, Vehtari A. Understanding predictive information criteria for Bayesian models. *Statistics and Computing* 2014, **24**:997–1016. doi: `https://doi.org/10.1007/s11222-013-9416-2`

[36] Vehtari A, Gelman A, Gabry J. Practical Bayesian model evaluation using leave one-out cross-validation and WAIC. *Statistics and Computing* 2017, **27**:1413–1432.

[37] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association* 2007; **102**(477):359–378. doi: `https://doi.org/10.1198/016214506000001437`

[38] Martínez-Bello D, López-Quílez A, Torres-Prieto A. Bayesian dynamic modeling of time series of dengue disease case counts. *PLoS Neglected Tropical Diseases* 2017; **11**(7):0005696.

[39] Reich BJ, Hodges JS, Zadnik V. Effect of residual smoothing on the posterior of the fixed effects in disease-mapping models. *Biometrics* 2006, **62**:1197–1206.

[40] Hodges JS, Reich BJ. Adding spatially-correlated errors can mess up the fixed effect you love. *The American Statistician* 2010, **64**:325–334.

[41] Fuentes-Vallejo M.: Space and space-time distributions of dengue in a hyper-endemic urban space: the case of Girardot, Colombia. *BMC Infectious Diseases* 2017; **17**:512

[42] Romero-Vega L, Pacheco O, de la Hoz-Restrepo F, Díaz-Quijano FA. Evaluation of dengue fever reports during an epidemic, Colombia. *Revista de Saude Publica* 2014; **48**(6):899–905

[43] Laughlin CA, Morens DM, Cassetti MC, Costero-Saint Denis A, San Martin JL, Whitehead SS, and Fauci AS. Dengue Research Opportunities in the Americas. *Journal of Infectious Diseases* 2012; **206**: 1121-7

Chapter 6

Spatio-temporal modeling of Zika and dengue infections within Colombia

Abstract

The aim of this study is to estimate the parallel relative risk of Zika virus disease (ZVD) and dengue using spatio-temporal interaction effects models, in one department and one city of Colombia during the 2015-2016 ZVD outbreak. We apply the integrated nested Laplace approximation (INLA) to parameter estimation, using the epidemiological week (EW) as a time measure. At departmental level, the best model showed that the dengue or ZVD risk in one municipality was highly associated with risk in the same municipality during the preceding EWs, while at city level, the final model selected established that the high risk of dengue or ZVD in one census sectors is highly associated not only with its neighboring census sectors in the same EW, but also with its neighboring sectors in the preceding EW. The spatio-temporal models provided risk smoothed estimates, risk credible intervals, estimating the probability of dengue and ZVD high risk by area and time period. We explore the intricacies of the modeling process and results interpretation, advocating for the use of spatio-temporal models of relative risk of dengue and ZVD in order to generate highly valuable epidemiological information for public health decision making.

Keywords: Disease mapping; Bayesian modeling, integrated nested Laplace approximation.

6.1 Introduction

Colombia is a country located in the north western corner of South America. Extensive regions of the country provide good conditions for the development of vector-borne diseases like dengue, malaria and yellow fever, among others [1]. In 2015 and 2016 , the Colombian population, like that of other South American countries [2, 3] was exposed to the Zika virus disease (ZVD), resulting in an estimated population of 106,659 people affected up to December 2016 [4]. During the same period, the incidence of dengue infections did not vanish but rather stayed at similar levels as in previous years [5].
In Colombia, the ZVD epidemics has been studied through descriptive epidemiological analyse [6–8], ecological analyses [9–13], and probabilistic modeling [14–16].
However, epidemiological disease mapping techniques, for example those used in dengue disease to characterize the spatial and temporal pattern in disease risk over extended geographical regions [17, 18], have not yet been applied in ZVD. Several levels of spatio-temporal modeling of dengue has been applied to continental, national, and municipal data in Brazil [19–22], Colombia [23–27], Ecuador [28, 29], and Indonesia [30]. Until recently most of the spatio-temporal relative risk models were developed under the Bayesian paradigm, applying Markov Chain Monte Carlo (MCMC) method. However, a new technique has emerged, providing support to modelers and epidemiologists to the use of the integrated Laplace approximation (INLA) [31], a fast and accurate tool for disease risk estimation.
Based on the modeling framework of spatio-temporal interaction effects models for relative risk developed by Knorr-Held [32], Martínez-Bello et al [26] applied spatio-temporal risk models including covariates in Colombia at the city level using MCMC, while Lowe et al [20–22], Wijayanti et al [30], and Abd Naeeim & Rahman [33] applied INLA to estimate relative risk in the development of large scale dengue warning systems in Brazil, Indonesia, and Malaysia, respectively.
The aim of this study is to model the spatio-temporal relative risk of ZVD and dengue in parallel during the period corresponding to the Zika outbreak in Colombia, from October 2015 to December 2016, in one high incidence city and one high incidence department of Colombia, using the epidemiological week (EW) as time measure and the census sector (city level) and municipality (departmental level) as geographic units. Our study has two specific objectives: first, to apply models describing the spatio-temporal risk of dengue and ZVD at two geographic aggregation levels, and second, to compare the risk of dengue and ZVD. For the first objective, the disease mapping models smoothed the risk of dengue and ZVD, improving the risk visualization by area and time period and generating credible intervals of risk. For the second objective, the model detected high risk areas of dengue and ZVD with a probability threshold defined by the data analyst. The probability estimates for relative risk permit the formulation of approximate hypotheses for detecting areas where the relative risk of dengue and ZVD is greater than one, with 95% probability.

6.2 Materials and Methods

6.2.1 Zika and dengue data in Santander and Bucaramanga, Colombia

Figure 6.1-a and Figure 6.1-b show the position of Colombia in the world and in the South American continent, respectively. Colombia (Figure 6.1-c) is divided into administratively autonomous *departments*, which in turn are divided into municipalities. Figure 6.1-d shows the department of Santander, located in north-eastern Colombia, covering an area of 30,537 km^2 and 2,071,016 people in 2016. Santander has 87 municipalities, and its administrative capital is the city of Bucaramanga (Figure 6.1-e), with an elevation of 959 meters, and an estimated population of 521,857 people spread over an urban area of 49 km^2 in 2016. Data on age, sex and address for people with the ZVD and dengue

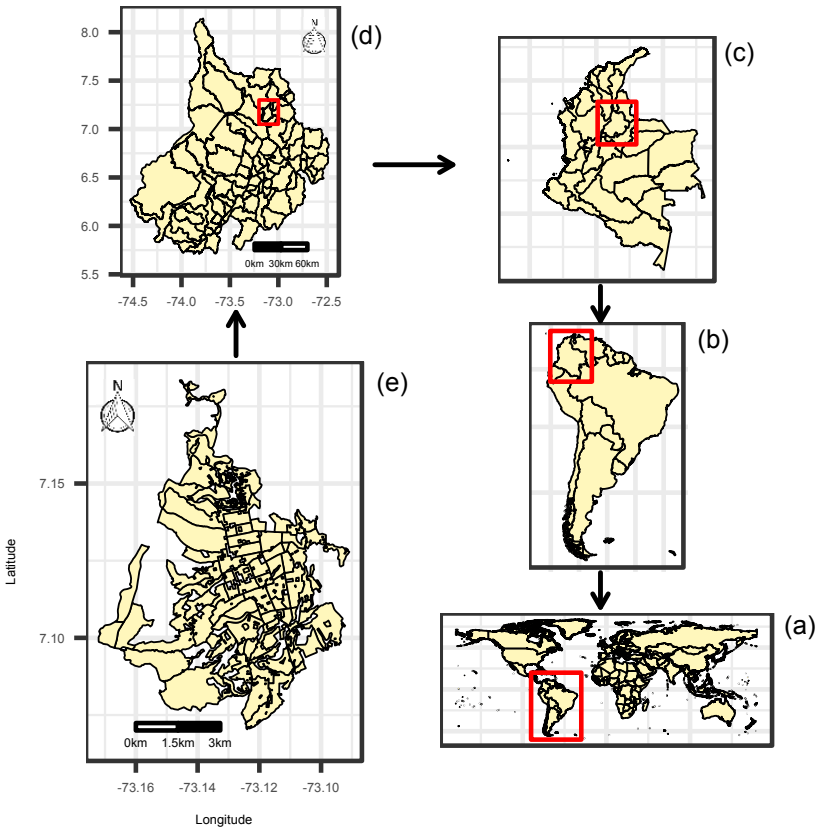


Figure 6.1: Geographical location of the study area: **(a)** world map; **(b)** South America; **(c)** Colombia; **(d)** department of Santander; **(e)** city of Bucaramanga

were obtained from the public health surveillance system (SIVIGILA) [34] for the period from the 42nd EW (October) of 2015 to the 52nd EW (December) of 2016. SIVIGILA is maintained by the Colombian National Health Institute (INS) and covers nearly 11,000

health services providers, plus 1117 municipalities, 32 departments and 5 districts. The objectives of SIVIGILA are to systematically collect, analyze, interpret, update, publish, and evaluate the data related with health events for the orientation of prevention and control activities in public health. SIVIGILA provides protocols for reporting each event of public health interest (EPHI) to be notified. In SIVIGILA, dengue, severe dengue and dengue mortality are registered with the codes 210, 220 and 580, respectively, while the ZVD code is 895. We considered the case definitions obtained from the SIVIGILA's protocols for dengue and ZVD. For dengue, we considered the cases that were classified as probable, confirmed by laboratory, and confirmed by epidemiological nexus, as well as the dengue mortality cases, while for ZVD, we considered the cases that were classified as probable, confirmed by clinical signs and confirmed by laboratory cases, as reported to the surveillance system weekly. At Santander's aggregation level, dengue and ZVD cases were aggregated by disease, municipality, and EW, and for the city of Bucaramanga, patients' addresses were geocoded using the ArcGIS Online Geocoding Service and then aggregated by disease, census sector, and EW. Census sectors are geographical areas employed by the Colombian statistical office to report census data. Census sectors correspond to the aggregation of 2 to 20 census blocks [35].

6.2.2 Expected values of ZVD and dengue

The next step was to compute ZVD and dengue expected values, which are necessary as input for the spatio-temporal risk models. Expected values were computed using the incidence rate (disease cases per 100,000 people) by five-year age group and sex at departmental and city level as follows:

$$\text{IR}_{kl} = \frac{\text{Cases}_{kl}}{\text{Population}_{kl}} \times 100,000$$

where IR_{kl} is the incidence rate in five-year age groups $k = 1,\ldots,K$, $(1 = [0,5), 2 = [5,10), \ldots, 14 = [65,100))$ and sex $l = 1,2$, (1=female, 2= male); Cases_{kl} are the total number of ZVD or dengue cases; and Population_{kl} is the total population in five-year age groups k and sex l for the departmental or city level over the complete study period, as obtained from the 2016 population of the Colombian census [36]. After computing the IR_{kl}, the expected values per small area (municipality for the departmental aggregation level, or census sector for the city aggregation level) were calculated as follows:

$$E_i = \sum_{k=1}^{K} \sum_{l=1}^{2} (\text{IR}_{kl} \times \text{Population}_{ikl})$$

where E_i is the expected value in small area i, and Population_{ikl} is the census population [36] in small area i, five-year age group k and sex l. Finally, the E_i are divided by the number of periods t, obtaining the expected values (E_{ij}) by area i and time period j. At the end, four sets of expected values were obtained: Santander's ZVD and dengue expected values, and Bucaramanga's ZVD and dengue expected values.

The ratio between the observed and the expected values of dengue and ZVD per area i and time period j is a statistic referred as the standardized incidence ratio (SIR). The SIR is a raw estimate of the disease risk, which can be modeled by the relative risk estimation assuming a probability distribution, producing point estimates and credible intervals for the risk, together with other valuable statistics, such as probability estimates of high risk per area and time period [37].

6.2.3 Spatio-temporal relative risk models

Let the observed counts of dengue or ZVD cases be O_{ij} where $i = 1,\ldots,n$ is the small area ($n = 87$ municipalities or $n = 94$ census sector) and $j = 1,\ldots,t$ ($t = 91$ EW) denotes the temporal unit. We assume that the observed counts are Poisson distributed with mean parameter μ_{ij} as follows:

$$\begin{aligned} O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\ \mu_{ij} &= E_{ij} r_{ij} \\ r_{ij} &= \exp(\eta_{ij}) \\ \eta_{ij} &= \alpha + \zeta_i + \gamma_j + \phi_j + \delta_{ij} \end{aligned} \tag{6.1}$$

where E_{ij} represents the expected values of dengue or ZVD calculated by internal or external standardization, and r_{ij} denotes the relative risk of dengue or ZVD by small area and EW, and η_{ij} is a linear predictor including latent variables accounting for the spatial, temporal and spatio-temporal dengue or ZVD risk structure. In Equation 6.1, α is a mean parameter, ζ_i accounts for the spatially structured risk pattern, γ_j and ϕ_j represent the temporally unstructured and structured risk pattern and δ_{ij} is the interaction term accounting for the spatio-temporal risk pattern. The probabilistic structure for the model parameters is

$$\begin{aligned} \boldsymbol{\zeta} &\sim \text{Normal}(\mathbf{0}, \sigma_\zeta^2 \mathbf{R}^{-1}) \\ \mathbf{R} &= (\rho_\zeta \mathbf{Q}_\zeta + (1-\rho_\zeta)\mathbf{I}_\zeta) \\ \boldsymbol{\gamma} &\sim \text{Normal}(\mathbf{0}, \sigma_\gamma^2 \mathbf{I}^{-1}) \\ \boldsymbol{\phi} &\sim \text{Normal}(\mathbf{0}, \sigma_\phi^2 \mathbf{Q}_\phi^{-1}) \end{aligned}$$

where the $\boldsymbol{\zeta}$ vector represents conditional autorregressive (CAR) Leroux spatially structured effects, representing the joint distribution of the structured spatial pattern, Normally distributed with zero mean vector, and variance-covariance matrix $\sigma_\zeta^2 \mathbf{R}^{-1}$. The $\mathbf{R}$ matrix follows Leroux et al [38], formulated by Ugarte et al [39, 40], where ρ_ζ is a smoothing parameter estimated from the data, $\mathbf{Q}_\zeta$ is an $n \times n$ proximity matrix, and $\mathbf{I}_\zeta$ is an $n \times n$ identity matrix. The $\mathbf{Q}_\zeta$ matrix contains ones where the area i is neighbor of area i', and zero otherwise. The $\boldsymbol{\gamma}$ vector represents the joint distribution of the unstructured temporal effects, Normally distributed with zero mean vector, and variance-covariance matrix $\sigma_\gamma^2 \mathbf{I}_\gamma^{-1}$, where σ_γ^2 is a variance parameter and $\mathbf{I}_\gamma$ is a $t \times t$ identity matrix. The $\boldsymbol{\phi}$ vector is the joint distribution of the structured temporal effects, Normally distributed

with zero mean vector, and variance-covariance matrix $\sigma_\phi^2 \mathbf{Q}_\phi^{-1}$, where σ_ϕ^2 is a variance parameter and $\mathbf{Q}_\phi$ is a $t \times t$ random walk 1 or 2 (RW1 or RW2) matrix. The joint distribution of the spatio-temporal interaction effects vector $\boldsymbol{\delta}' = [\delta_{11}, \delta_{12} \ldots, \delta_{nt}]$ is Normally distributed with mean zero vector and variance-covariance matrix defined by one out of four interaction types as follows:

$$\boldsymbol{\delta} \sim \text{Normal}(\mathbf{0}, \sigma_\delta^2 (\mathbf{I}_\gamma \otimes \mathbf{I}_\zeta)^{-1}) \tag{6.2}$$

$$\boldsymbol{\delta} \sim \text{Normal}(\mathbf{0}, \sigma_\delta^2 (\mathbf{Q}_\phi \otimes \mathbf{I}_\zeta)^{-1}) \tag{6.3}$$

$$\boldsymbol{\delta} \sim \text{Normal}(\mathbf{0}, \sigma_\delta^2 (\mathbf{Q}_\zeta \otimes \mathbf{I}_\phi)^{-1}) \tag{6.4}$$

$$\boldsymbol{\delta} \sim \text{Normal}(\mathbf{0}, \sigma_\delta^2 (\mathbf{Q}_\phi \otimes \mathbf{Q}_\zeta)^{-1}) \tag{6.5}$$

here for all types of interaction effects, σ_δ^2 is a variance parameter and $\otimes$ is the Kronecker product of two matrices. We follow the Knorr-Held [32] interaction effects taxonomy. Equation 6.2 shows the type I interaction effects $\boldsymbol{\delta}$ with structure matrix $(\mathbf{I}_\gamma \otimes \mathbf{I}_\zeta)$, representing an unstructured interaction effect, where $\mathbf{I}_\gamma$ is a $t \times t$ identity matrix and $\mathbf{I}_\zeta$ is a $n \times n$ identity matrix. Equation 6.3 displays the type II interaction effects $\boldsymbol{\delta}$ with structure matrix $(\mathbf{Q}_\phi \otimes \mathbf{I}_\zeta)$ modeling a temporal interaction effect, where $\mathbf{Q}_\phi$ is $t \times t$ RW1 or RW2 matrix, and $\mathbf{I}_\zeta$ defined above. Equation 6.4 shows the type III interaction effects $\boldsymbol{\delta}$ with structure matrix $(\mathbf{Q}_\zeta \otimes \mathbf{I}_\gamma)$ representing spatial interaction effects, where $\mathbf{Q}_\zeta$ is an $n \times n$ proximity matrix defined above, and $\mathbf{I}_\gamma$ is a $t \times t$ matrix previously defined. Finally, Equation 6.5 shows the type IV interaction effects $\boldsymbol{\delta}$ with structure matrix $(\mathbf{Q}_\phi \otimes \mathbf{Q}_\zeta)$, defining inseparable spatio-temporal interaction effects, where $\mathbf{Q}_\phi$ is a $t \times t$ RW1 or RW2 matrix and $\mathbf{Q}_\zeta$ is a $n \times n$ proximity matrix.

6.2.4 Inference

The Bayesian inference for the spatio-temporal models is currently developed applying MCMC methods, through conditional probability distributions of the model parameters; however, we followed the INLA [31] technique to fit the spatio-temporal interaction effects models to the data, using the `INLA` package downloaded from `www.r-inla.org` and the `R` software version 3.3 [41]. We follow Goicoa et al's [42] constraints specification to address the parameter identifiability issues in the interaction terms. The hyper-prior specification for the mean parameter (α), the variance hyper-parameters (σ_ζ^2, σ_γ^2, σ_ϕ^2 and σ_δ^2) and the smoothing parameter (ρ_ζ) follows

$$\alpha \sim \text{Normal}(0, 1000)$$

$$\rho_\zeta \sim \text{Uniform}(0, 1)$$

$$\sigma_\zeta,\ \sigma_\gamma,\ \sigma_\phi,\ \sigma_\delta \sim \text{Uniform}(0, \infty)$$

Model selection was based on the analysis of the deviance, the effective number of parameters,the Watanabe-Akaike information criterion (WAIC) by Watanabe [43] and the logarithmic score (LS) by Gneiting and Raftery [44] implemented by Ugarte et al [40]. The selected final model corresponds to the model displaying the lowest WAIC and LS.

6.3 Results

6.3.1 Exploratory data analysis

First, we obtained from the Santander's SIVIGILA database a total of 10,051 ZVD cases (63.1% females and 36.9% males) and 7891 dengue cases (48.6% females and 51.4% males), while Bucaramanga's database included 3662 ZVD cases (61.2% females and 38.8% males) and 2470 dengue cases (49.3% females and 50.7% males).
Then, using the number of dengue and ZVD cases by age group and sex, along with the population of the department of Santander and the city of Bucaramanga, we calculated the incidence rate (cases by 100,000 people) by age and sex. Figure 6.2 displays the incidence rate by ten-year age group and sex for dengue and ZVD at departmental (Santander) and city level (Bucaramanga). In general, incidence rates were higher at the city level than in the department as a whole. Both geographic levels presented a similar incidence pattern for dengue and ZVD. Dengue disease incidence is slightly higher in younger than older people, but it is very similar for women and men, while ZVD incidence was higher in the age range of 20 to 65 years for women than men. The incidence rates were used to generate the expected values per area and time period, therefore, the study uses internal standardization. Thereafter, the expected values for dengue and ZVD at departmental and city level per area and EW were combined with the observed values to produce the standardized incidence ratio (SIR) values per area and time period. Figure 6.3 presents

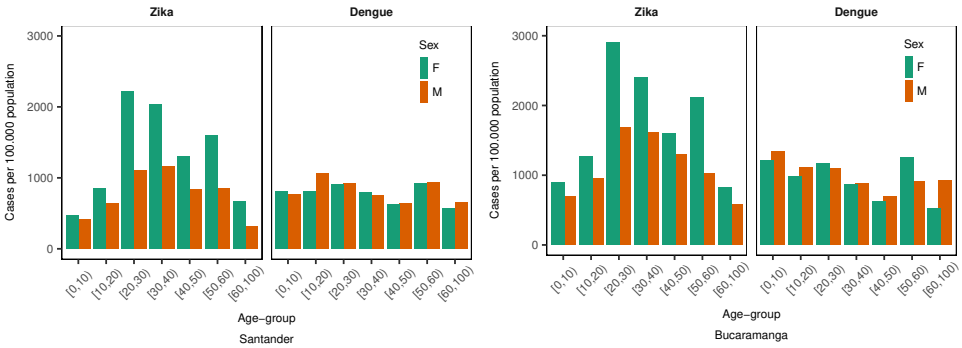


Figure 6.2: Incidence rate of ZVD and dengue (cases per 100,000 people) by ten-year age group and sex in the department of Santander and the city of Bucaramanga, October 2015 - December 2016

the longitudinal profiles of the SIR by aggregation area for dengue and ZVD for the study period. For the department of Santander, the ZVD SIR profiles show values of more than 40 units in two municipalities, but in general most of the municipalities did not surpass SIR values of 20 units. For dengue, the SIR is consistently under 20 units throughout the study period in most municipalities, with two spikes in incidence in December 2015 and May 2016. For the city of Bucaramanga, the SIR for ZVD was under 20 units during the outbreak between January 2016 and July 2016, and for dengue there was a constant SIR

of less than 20 units during the study period.

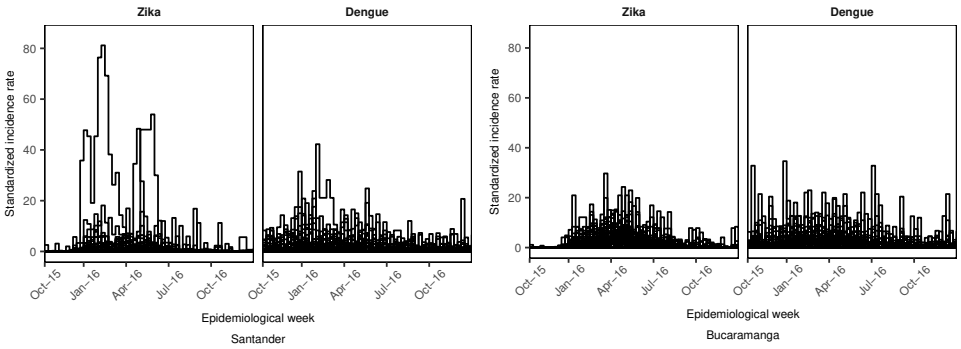


Figure 6.3: Longitudinal profiles of the standardized incidence ratio (SIR) of dengue and ZVD by municipality (department of Santander) and census sector (city of Bucaramanga), October 2015- December 2016

SIR can also be accumulated in the study period and mapped by area employing a choropleth map to visualize SIR patterns. Figure 6.4 shows maps of the accumulated SIR for dengue and ZVD by aggregation area in Santander and Bucaramanga for the study period. The accumulated ZVD SIR in Santander shows few municipalities with high SIR levels in the northern, central and eastern areas of the department, while high values of the accumulated dengue SIR are presented to the southern municipalities of the department. In contrast, at the city aggregation level, high values of the accumulated ZVD and dengue SIR are apparent in the central census sectors of the city, displaying similar distribution patterns for both diseases.

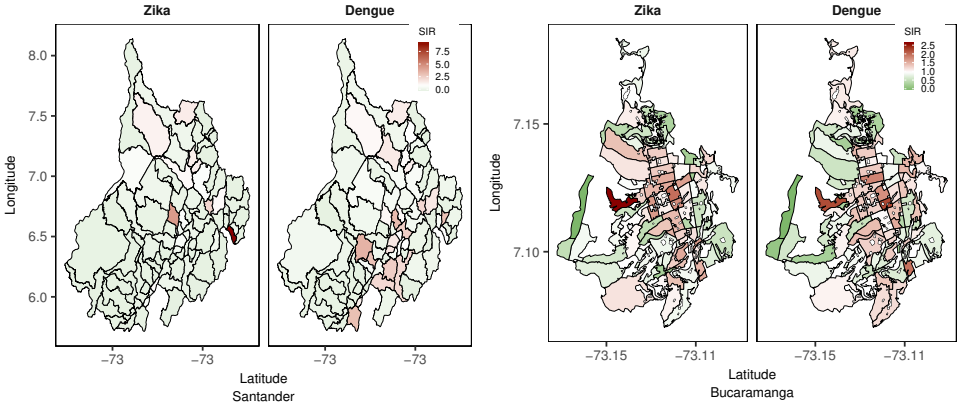


Figure 6.4: Accumulated standardized incidence ratio (SIR) maps of dengue and ZVD in Santander and Bucaramanga, October 2015 - December 2016.

6.3.2 Model findings

Table 6.1 displays the selection criteria statistics for the spatio-temporal relative risk models of dengue and ZVD for Santander and Bucaramanga. At the departmental aggregation level, and using the WAIC and the LS, the selected model contains Type II interaction effects (temporal effects) for both dengue (WAIC = 7413.6 and LS = 3738.7) and ZVD (WAIC = 4562.9 and LS = 2362.7). At the city aggregation level, the best model corresponds to the one with Type IV interaction effects (spatio-temporal inseparable effects) for dengue (WAIC = 8010.4 and LS = 4005.7) and ZVD (WAIC = 7804.1 and LS = 3904.1). Table 6.2 presents the smoothing parameters (ρ_ζ) of the spatially

Table 6.1: Selection statistics from the spatio-temporal interaction effects models of ZVD and dengue $\eta_{ij} = \alpha + \zeta_i + \gamma_j + \phi_j + \delta_{ij}$, and RW1 temporally structured effects, for the department of Santander and the city of Bucaramanga, October 2015 - December 2016. Effective number of parameters p_{eff}, Watanabe Akaike information criterion (WAIC), and logarithmic score (LS).

	Zika				**Dengue**			
	Deviance	**p_{eff}**	**WAIC**	**LS**	**Deviance**	**p_{eff}**	**WAIC**	**LS**
				Department of Santander				
No interaction	6290.6	115.7	6592.6	3307.4	8308.4	112.7	8490.5	4236.5
Type I	4222.8	630.5	4855.0	4570.9	7278.4	623.9	7934.2	4156.7
Type II	4166.2	410.4	4562.9	2362.7	6979.7	435.3	7413.6	3738.7
Type III	4247.6	589.1	4856.3	4147.4	7314.4	589.0	7968.7	4143.6
Type IV	4206.4	382.1	4593.7	2437.6	7044.8	408.1	7470.9	3764.0
				City of Bucaramanga				
No interaction	7852.1	105.0	7970.1	3985.3	7938.5	93.9	8042.5	4021.4
Type I	7659.3	275.0	7949.3	3980.8	7680.3	318.2	8018.5	4014.6
Type II	7537.5	263.7	7808.2	3907.3	7841.2	165.6	8024.2	4012.8
Type III	7653.2	247.7	7913.6	3961.5	7865.6	159.9	8040.7	4021.2
Type IV	7576.6	218.7	7804.1	3904.1	7839.7	155.8	8010.4	4005.7

structured random effects and the standard deviation hyperparameters corresponding to the final selected models in Santander and Bucaramanga. At departmental level, the smoothing parameters display posterior mean values of 0.61 and 0.50 for ZVD and dengue respectively, which denotes moderate spatially structured effects. We observe similar results at city level, where the posterior means of the smoothing parameters were 0.55 and 0.49 for ZVD and dengue, respectively. The standard deviation hyperparameters at departmental level are higher for ZVD than for dengue, and the small standard deviation of the temporally structured effects (σ_ϕ) denotes the small variability of the $\boldsymbol{\phi}$. At city level, the (σ_ϕ) also shows the small variability for all the hyperparameters. Using the selected models for dengue and ZVD in Santander (model with Type II interaction effects) and Bucaramanga (model with Type IV interaction effects), the probability of the structured spatial risk pattern greater than 1 given the observed disease counts ($P(\exp(\zeta_i > 1|\mathbf{O})$) of the entire study period was mapped in the Figure 6.5. What can this probability tell to

Table 6.2: Standard deviations hyper-parameters of the interaction effects models for the relative risk of ZVD and Dengue, October 2015 - December 2016. Posterior mean, standar deviation (SD), and 2.5%, 50%, and 97.5% percentiles.

	Zika					**Dengue**				
	Mean	**SD**	**2.5%**	**50%**	**97.5%**	**Mean**	**SD**	**2.5%**	**50%**	**97.5%**
Department of Santander										
ρ_ζ	0.61	0.17	0.26	0.63	0.90	0.50	0.18	0.17	0.50	0.84
σ_ζ	2.36	0.33	1.80	2.34	3.05	2.04	0.32	1.49	2.02	2.73
σ_γ	0.32	0.04	0.24	0.31	0.41	0.11	0.02	0.08	0.11	0.16
σ_ϕ	0.06	0.04	0.01	0.05	0.16	0.04	0.02	0.01	0.04	0.09
σ_δ	0.36	0.02	0.31	0.36	0.40	0.23	0.01	0.20	0.23	0.25
City of Bucaramanga										
ρ_ζ	0.55	0.20	0.16	0.55	0.89	0.49	0.19	0.15	0.49	0.84
σ_ζ	0.48	0.07	0.36	0.48	0.64	0.52	0.08	0.39	0.51	0.68
σ_γ	0.42	0.06	0.33	0.42	0.54	0.19	0.04	0.13	0.18	0.27
σ_ϕ	0.07	0.06	0.01	0.06	0.25	0.07	0.04	0.02	0.07	0.16
σ_δ	0.17	0.02	0.14	0.17	0.21	0.10	0.02	0.07	0.10	0.14

the epidemiologists, biostatisticians and public health officials? The mapped probability provides estimates of the areas at departmental and city level where the affected areas share high risk of dengue or ZVD. From Figure 6.5 in the department of Santander, the red areas reveal the municipalities with probabilities close to 1, displaying high risk with a clustered pattern, so for ZVD, the northern and western municipalities present the spatial clustered high-risk, while for dengue, the north and south municipalities show the spatial high-risk pattern. At city level, the probability close to 1 for the distribution of high-risk spatial clusters of census sectors is similar for both dengue and ZVD, and concentrated in the central parts of the city, following the pattern displayed by the SIR maps in Figure 6.4. In addition to the spatial structured effects, the relative risk per area and time period can be obtained from the final selected spatio-temporal models. For the department of Santander, Figure 6.6 shows the longitudinal profiles of relative risk of dengue (gray) and ZVD (pink) displaying the posterior mean and 95% credible intervals for selected municipalities. Before examining the plot, we warn the reader that different scales for the relative risk are displayed for the selected municipalities, because the high variability of the risk profiles from municipality to municipality would make it difficult to visualize the trend in several areas. The municipalities were selected based on the high probability of spatial structured pattern from Figure 6.5. Three nearby municipalities (Bucaramanga, Girón, and Florida) reveal similar risk patterns for both diseases, showing the ZVD outbreak closely following the dengue outbreak in the first semester of 2016, while municipalities such as Cimitarra and Rionegro showed low risk for dengue and ZVD, and Capitanejo displayed the highest relative risk of all the

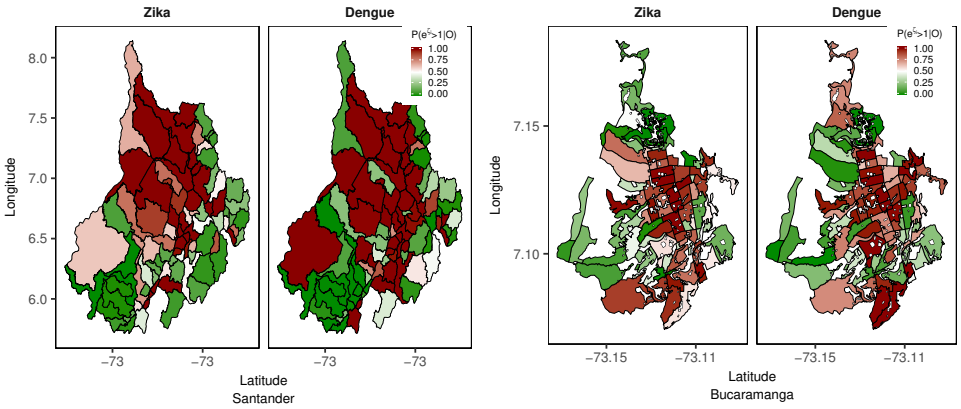


Figure 6.5: Probability of spatial structured effects greater than 1 $[P(\exp(\zeta_i) > 1|\mathbf{O})]$, department of Santander and city of Bucaramanga, October 2015 - December 2016.

municipalities in the department. At the city level (Bucaramanga), some selected census sectors with probability close to 1 for the spatial structured effects did not reveal the high variability presented at departmental level, showing the high risk of ZVD in close connection with that of dengue, and with the ZVD risk being higher than the dengue risk for the selected census sectors. In the search for a combined risk representation of

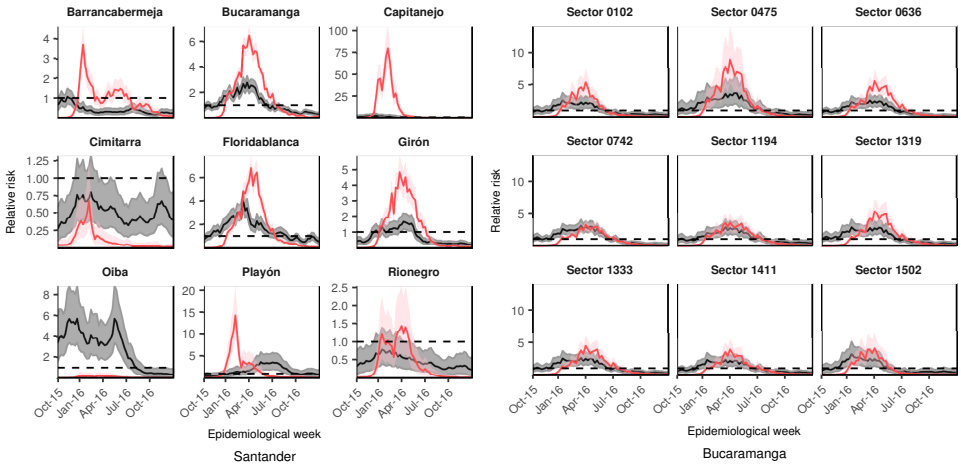


Figure 6.6: Selected longitudinal profiles of the posterior means and 95% credible intervals for the relative risk of dengue (gray shadow) and ZVD (pink shadow), department of Santander and city of Bucaramanga, October 2015 - December 2016. Dashed lines correspond to relative risk equal to 1.

dengue and ZVD, Figure 6.7 explores the spatio-temporal representation of relative risk greater of 1 with 95% probability for both diseases, that is, the areas where the 95% credible intervals of the risk for dengue and ZVD are higher than 1. In that order, the

heatmap for Santander represents the municipalities with risk greater than 1 with 95% probability, following the selected model for dengue and ZVD in Santander, and exposing the strictly temporal risk pattern by municipality, with few municipalities sharing relative risk greater than 1 with 95% probability. The shared relative risk greater than 1 with 95% probability displays quite a different pattern at municipality level. For the ZVD outbreak period (January to July 2016), almost all the census sectors demonstrated relative risk greater than 1 with 95% probability for both diseases, taking into account that the risk representation follows the selected final model for dengue and ZVD corresponding to the Type IV interaction effects model (spatio-temporal inseparable interaction). Not only can

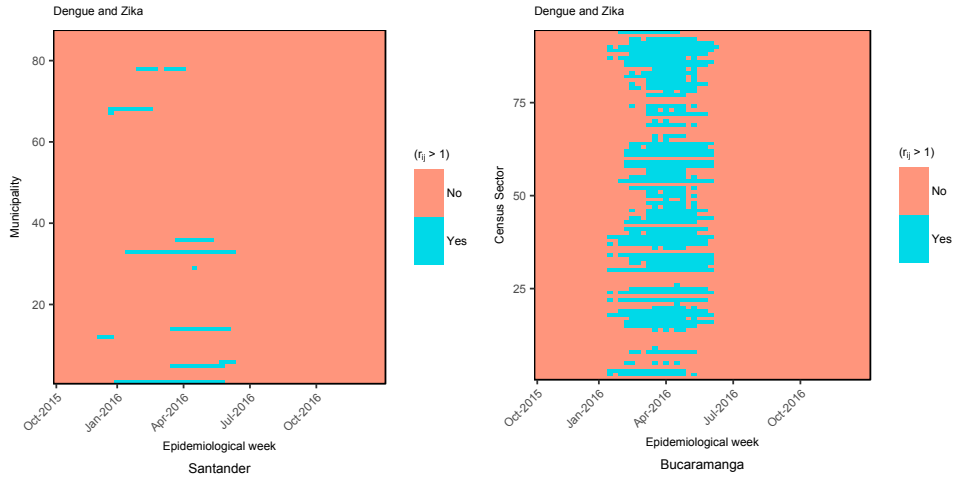


Figure 6.7: Heatmap of the relative risk greater than 1 ($r_{ij} > 1$) with 95% probability for dengue and ZVD in the department of Santander and the city of Bucaramanga, October 2015 - December 2016.

the relative risk be presented as in Figure 6.6 using longitudinal posterior mean profiles and credible intervals, but it can also be mapped directly (posterior mean, or selected percentiles) or represented by the probability of a relative risk greater than 1 given the observed cases counts ($P(r_{ij} > 1|\mathbf{O})$). Figure 6.8 exhibits maps of the probability of relative risk greater than 1 for dengue and ZVD in Santander and Bucaramanga, from the fifth to the eighth EWs of 2016 (the period associated with the rapid dissemination of ZVD at departmental and city level). For those EWs, municipalities with high probability of relative risk greater than 1 for dengue in Santander are revealed to the central northern and southern municipalities, while for ZVD, high-probability municipalities are in the northwestern region of the department. In Bucaramanga, almost all census sectors display high probability of relative risk greater than 1 for dengue, while, many census sectors also revealed high probability for ZVD during the same time period across the city.

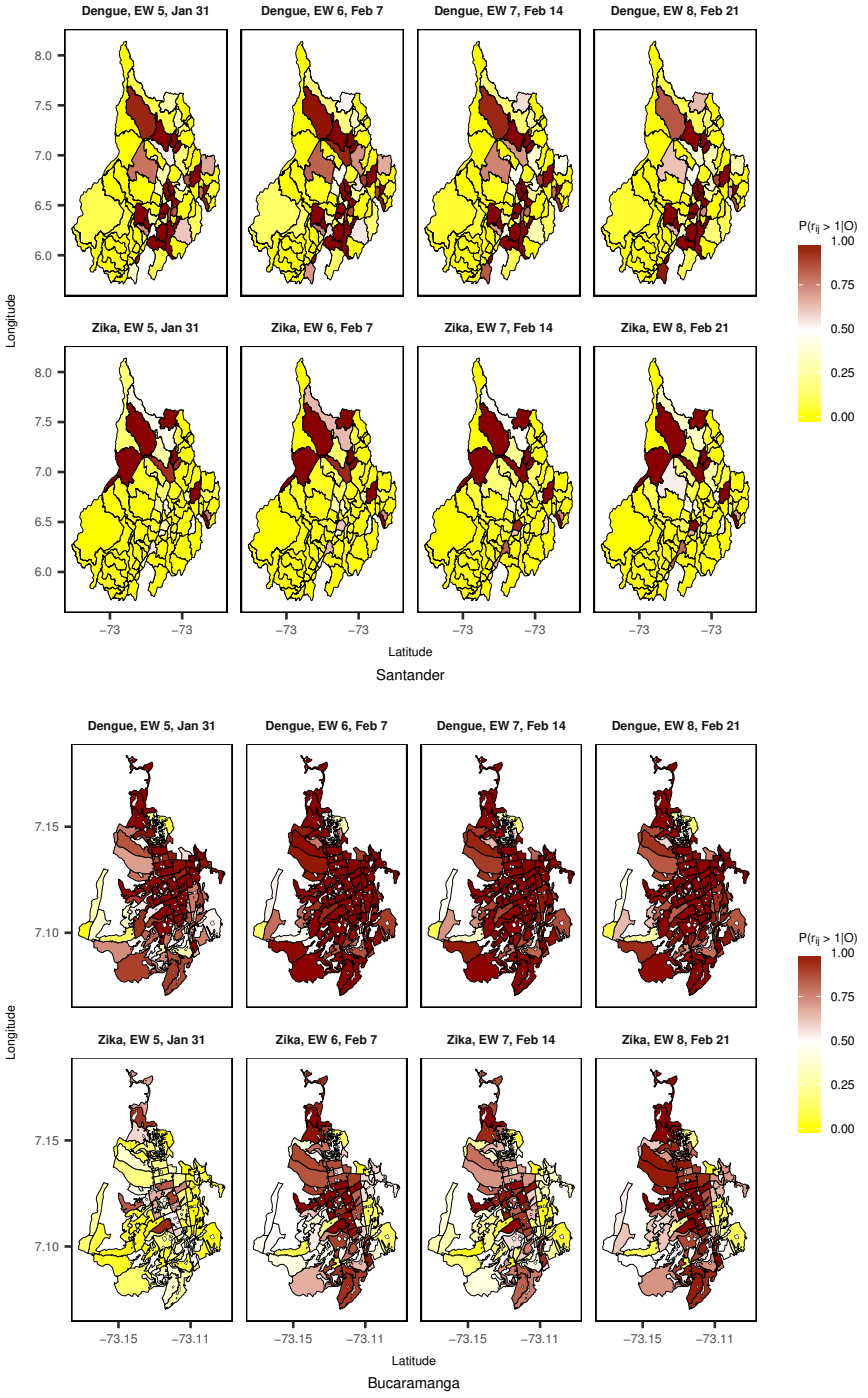


Figure 6.8: Evolution of the probability of dengue and ZVD relative risk greater than 1 given the observed cases ($P(r_{ij} > 1|\mathbf{O})$), for selected EW in January and February 2016, in the department of Santander and the city of Bucaramanga.

6.4 Discussion

The present paper applies parallel spatio-temporal interaction effect models of relative risk to ZVD and dengue data at two geographic levels of aggregation, first at departmental level (municipalities aggregation), and second, at city level (census sector aggregation). The aim is to provide risk estimates for arboviral diseases elucidating the risk transmission dynamic of dengue and ZVD.
We fitted the spatio-temporal models using the integrated nested Laplace Approximations (INLA), which is a fast and accurate numerical method to develop Bayesian analysis.
The within-sample predictive measures WAIC and LS were the information criteria applied to select the final models at departmental and city level. The model selection was based on the models showing the lowest value for the information criteria. At departmental level, the final selected model included Type II (temporal interaction) spatio-temporal interactions effects, CAR-Leroux structured spatial effects and RW1 structured temporal effects. Selecting the Type II model for inferences implies that the dengue or ZVD risk in one municipality is highly associated with risk in the same municipality during the preceding EW. At city level, the final selected model included Type IV (inseparable) spatio-temporal interaction effects, CAR-Leroux structured spatial effects and RW1 structured temporal effects. The election of the Type IV model implies that the risk of dengue or ZVD in one census sectors is highly associated not only with its neighboring census sectors in the same EW, but also with its neighboring sectors in the preceding EW.
The information concerning the relative risk obtained from the models can be presented in many ways. We provide longitudinal profiles of relative risk (posterior mean and 95% credible intervals) and maps of the high-risk probability by municipality or census sector. In addition, we represented the joint risk of dengue and ZVD using 95% credible intervals, by means of creating a categorical variable representing high risk when the lower limit of the 95% credible interval was greater than 1, identifying municipalities or census sectors at high risk with 95% probability for dengue and ZVD. At departmental level we distinguished 11 municipalities displaying high risk for dengue and ZVD, while at the city level, almost all the census sectors showed joint high risk for both diseases.
It is difficult to compare the findings in our study with other results in the dengue and ZVD literature, because a parallel risk assessment of ZVD and dengue under a spatio-temporal setting is not a current research practice, perhaps due to the recent emergence in ZVD in the affected countries. For the Colombian case, Krystosik et al. [45] associate perceived risk factors such as proximity to standing water, canals, poverty, localized violence, and military migration with density maps of dengue, ZVD and Chikungunya virus disease (CVD) cases at city level, using spatial point pattern methods. Krystosik et al. identified suitable areas for disease transmission, ignoring the temporal nature of the data, which our study considers, although we did not consider an association between covariates for environmental or socio-economic factors and dengue or ZVD.
Chien et al [46] published another case study using Colombian data, exploring a quasi-Poisson spatial nonlinear model to investigate the association among weekly ZVD infection and lagged effects of weather variables, using the department as spatial unit, from

January 2015 to December 2016. The modeling results suggest that mean humidity, total rainfall, and the maximum temperature can help predict a ZVD infection outbreak at least 3 months in advance, postulating the model as applicable in the development of a ZVD surveillance system by public health authorities. Chien et al's study also modeled the relative risk of ZVD; however, the relative risk was based on non-smoothed estimates. In addition, they ignored the spatio-temporal nature of the outcome variable while investigating the lagged effects of the covariates, so their results were based on a fixed-in-time outcome variable and ignored the risk association between areas mediated by the spatial nature of the data. They also used a spatial aggregation unit much larger than that presented in our study, and finally, they analyzed only ZVD, not dengue or CVD.

For the city of Bahía de Caráquez in Ecuador, (a country neighboring Colombia), Stewart-Ibarra et al [47] associated the degree of psychological distress with the presence of dengue, CVD, and Zika infections in a suburban community after the 2016 earthquake. The study followed a survey methodology, without considering the temporal or spatial nature of the data and collapsing all three arboviral diseases studied into one composite variable. Thus, this paper does not investigate the possible differences in transmission for the diseases.

In the course of the concurrent disease outbreaks of CVD and ZVD in French Polynesia and the French West Indies between 2013 and 2016, Riou et al. [48] modeled the joint transmission of both diseases using a Bayesian time-dependent susceptible-infectious-recovered (TSIR) framework [49], including information on weather conditions (temperature and precipitation). This study is interesting because the findings can be related to our results. The authors found that after controlling for locality and weather conditions, the difference in transmission between the two viruses was minimal, suggesting that the epidemic dynamics are more dependent on factors related to the mosquito abundance and local environmental and socio economic conditions than due to the intrinsic characteristics of the viruses. In addition, the authors modelled the risk between ZVD and CVD, considering the between-island association, which we also accounted in our modelling framework.

In Brazil, one of the American countries most affected by the 2015-2016 ZVD epidemic, Aguiar et al. [50] assessed the potential spatial risk of ZVD and CVD using a maximum entropy (MaxEnt) approach, estimating the optimal conditions for infection, based on the disease occurrence and environmental and social factors for 2015-2016 at municipality level. A local empirical Bayesian (LEB) analysis was performed using the cases of ZVD and CVD reported by municipalities, and then, the municipalities with LEB rates greater than 0 were included into the MaxEnt analysis. The analyses provided joint risk maps of ZVD and CVD. Although Aguiar et al's study included covariates for the analysis, they also ignored the spatio-temporal nature of the data, such as those described in the previous studies shown in the current section.

The spatio-temporal models of relative risk shown here are available to study dengue and ZVD transmission, acting as predictive models to help public health officers to anticipate the time and place where the next ZVD or other emerging arboviral disease outbreak will occur [51]. Identifying high risk areas of vector borne diseases at small geographic

resolution using within-sample predictive models or short-time-ahead, out-of-sample predictive models supports control and prevention of vector-transmitted diseases, which is one the advantages of the spatio-temporal methods applied in the present study [52]. The modeling approach in our study could easily be combined with covariates originating from vectorial ecology, socio economic and infrastructural data, improving the information generated from the modeling exercise, as demonstrated for dengue by Wijayanti et al. [30]. Such studies could clarify the role of factors such as land cover, meteorological, and socio-economic determinants of arboviral diseases [53].
This study presents some limitations. For the dengue data, it is well known that underreporting could reach almost 70%, because many of the affected people do not approach health services, while some of the patients attending the health services are not correctly diagnosed [54]. There are few ZVD studies addressing the quality of the reporting and diagnosis; however, one study from the National Health Institute of Colombia, carried out in a hyper-endemic ZVD town in Colombia, described around 49% of ZVD underreporting for people detected in a community active search of ZVD cases, and 81% of ZVD underreporting for people detected in a health service active search [55]. The intrinsic characteristics of the data make the results of the study prone to the same inaccuracies as the event reports produced by the public health surveillance system, because we used exactly the same data registered and analyzed by the system. For instance, the system's event notification depends on the medical doctor's ability to recognize and register the notifiable event, only using laboratory testing for severe dengue cases, or pregnant women and children for ZVD. Even with the surveillance system underreporting ZVD and dengue, the spatio-temporal risk estimates would complement the current information generated by the system.
To end our discussion, we need to mention that we modeled ZVD and dengue independently before combining the risk to obtain a joint risk estimation. However, a framework of spatial joint models of relative risk exist [37] and could be explored to include several diseases at the same time. In addition, it would be possible to fit longitudinal spatio-temporal joint models, providing a full structure for estimation of relative risk, although the implementation is not straightforward. A following step of our research is to test joint spatial and spatio-temporal models of dengue, ZVD, CVD and other emerging vector-borne infectious diseases.

6.5 Conclusion

We have explored a statistical method to express the risk of ZVD and dengue using spatio-temporal models of relative risk, providing a coherent, systematic and structured approach to model implementation, applying the INLA modern technique for Bayesian parameter estimation. We argue that the spatio-temporal risk framework has advantages for the public health officers, epidemiologists, biostatisticians, and data analysts involved in controlling vector-borne infectious diseases, corresponding with the fast and accurate

risk estimation (provided that good quality data are available), the risk smoothing for trend visualization, and the generation of credible intervals expressing the risk probability, which could be portrayed in spatial or temporal features. We also observed shared areas of ZVD and dengue high risk with high probability, given a probability threshold imposed by the data analyst. We have searched the literature for parallel modeling of dengue and ZVD risk, detecting a lack of spatio-temporal risk models for arboviral diseases. This study thus presents a helpful statistical method as a tool for decision making in preventing and controlling vector-borne emergent diseases.

6.6 References

[1] Villar LA, Rojas DP, Besada-Lombana S, Sarti E. Epidemiological Trends of Dengue Disease in Colombia (2000-2011): A Systematic Review. *PLoS Neglected Tropical Diseases.* 2015, **9**, 1–16.

[2] Plourde AR, Bloch EM. A Literature Review of Zika Virus. *Emerging Infectious Diseases.* 2016, **22**, 1185-1192.

[3] Pan American Health Organization/World Health Organization.*Zika suspected and confirmed cases reported by countries and territories in the Americas Cumulative cases, 2015-2017.* Updated as of 5 January 2017. Washington, D.C.: PAHO/WHO, 2017, Pan American Health Organization. URL: www.paho.org

[4] National Health Institute. *Weekly epidemiologic bulletin: epidemiological week number 52 of 2016, 25 December - 31 December.* National Health Institute, Direction of Surveillance and Risk Analysis in Public Health. Bogotá D.C., Colombia. [In spanish]. URL: `https://www.ins.gov.co/buscador-eventos/BoletinEpidemiologico/2016%20Bolet%C3%ADn%20epidemiol%C3%B3gico%20semana%2052%20-.pdf`

[5] National Health Institute. *Event report: dengue, 2016*. National Health Institute, Direction of Surveillance and Risk Analysis in Public Health. Bogotá D.C., Colombia. [In spanish]. URL: `https://www.ins.gov.co/buscador-eventos/Informesdeevento/Dengue%202016.pdf`

[6] Pacheco O, Beltrán M, Nelson CA et al. Zika Virus Disease in Colombia - Preliminary Report. *New England Journal of Medicine.* 2016. DOI: 10.1056/NEJMoa1604037

[7] Rojas DP, Dean NE, Yang Y, Kenah E, Quintero J, Tomasi S, Ramirez EL, Kelly Y, Castro C, Carrasquilla G, Halloran ME, Longini IM. The epidemiology and transmissibility of Zika virus in Girardot and San Andres island, Colombia, September 2015 to January 2016. *Euro Surveillance.* 2016, **21**, 30283.

[8] Tolosa N, Tinker SC, Pacheco O, Valencia D, Salas-Botero D, et al. Zika Virus Disease in Children in Colombia, August 2015 to May 2016 *Paediatric and Perinatal Epidemiology.* 2017, **31**, 537-545.

[9] Rodriguez-Morales AJ, Patiño-Cadavid LJ, Lozada-Riasco CO, Villamil-Gómez WE. Mapping Zika in municipalities of one coastal department of Colombia (Sucre) using geographic information systems during the 2015–2016 outbreak: implications for public health and travel advice. *International Journal of Infectious Diseases.* 2016, **48**, 70-72.

[10] Rodriguez-Morales AJ, Galindo-Marquez ML, García-Loaiza CJ et al. Mapping Zika virus infection using geographical information systems in Tolima, Colombia, 2015-2016 [version 1; referees: 2 approved] *F1000Research.* 2016, **5**, 568.

[11] Rodriguez-Morales AJ, García-Loaiza CJ, Galindo-Marquez ML, Sabogal-Roman JA, Marin-Loaiza S, Lozada-Riascos CO, Díaz-Quijano FA. Zika infection GIS-based mapping suggest high transmission activity in the border area of La Guajira, Colombia, a northeastern coast Caribbean department, 2015-2016: Implications for public health, migration and travel. *Travel Medicine and Infectious Diseases.* 2016, **14**, 286-288.

[12] Rodriguez-Morales AJ, Haque U, Ball JD, García-Loaiza CJ, Galindo-Marquez ML, Sabogal-Romanm JA, Marin-Loaiza S, Ayala AF, Lozada-Riascos CO, Diaz-Quijano FA, Alvarado-Socarras JL. Spatial distribution of Zika virus infection in Northeastern Colombia. *Infezioni in Medicina.* 2017, **3**, 241-246.

[13] Rodriguez-Morales AJ, Ruiz P, Tabares J, Ossa CA, Yepes-Echeverry MC, Ramirez-Jaramillo V, Galindo-Marquez ML, García-Loaiza CJ, Sabogal-Roman JA, Parra-Valencia E, Lagos-Grisales GJ, Lozada-Riascos CO, de Pijper CA, Grobusch M. Mapping the ecoepidemiology of Zika virus infection in urban and rural areas of Pereira, Risaralda, Colombia, 2015-2016: Implications for public health and travel medicine. *Travel Medicine and Infectious Diseases.* 2017, **18**, 57-66.

[14] Towers S, Brauer F, Castillo-Chavez C, Falconar AKI, Mubayi A, Romero-Vivas CME. Estimate of the reproduction number of the 2015 Zika virus outbreak in Barranquilla, Colombia, and estimation of the relative role of sexual transmission. *Epidemics.* 2016, **17**, 50-55.

[15] Ospina J, Hincapie-Palacio D, Ochoa J, Molina A, Rua G, Pajaro D, Arrubla M, Almanza R, Paredes M and Mubayi A. Stratifying the potential local transmission of Zika in municipalities of Antioquia, Colombia. *Tropical Medicine and International Health.* 2017, **22**, 1249-1265.

[16] Sebrango-Rodríguez CR, Martínez-Bello DA, Sánchez-Valdés L, Thilakarathne PJ, Del Fava E, Vand Der Stuyft P, López-Quílez A, Shkedy Z. Parameter Estimation

and Real-Time Predictions of a Single Zika Outbreak Using Model Averaging. *Epidemiology and Infection.* 2017, **145**, 2313-2323.

[17] Lee D. A comparison of conditional autoregressive models used in Bayesian disease mapping. *Spatial and Spatio-temporal Epidemiology.* 2011, **2**, 79–89.

[18] Louis VR, Phalkey R, Horstick O, Ratanawong P, Wilder-Smith A, Tozan Y, Dambach P. Modeling tools for dengue risk mapping - a systematic review. *International Journal of Health Geographics.* 2014, **13**, 50.

[19] Lowe R, Bailey TC, Stephenson DB, Graham RJ, Coelho CAS, Carvalho M, Barcellos C. Spatio-temporal modelling of climate-sensitive disease risk: Towards an early warning system for dengue in Brazil. *Computers in Geosciences.* 2011, **37**, 371–381.

[20] Lowe R, Bailey T, Stephenson D, Jupp T, Graham R, Barcellos C, Carvalho M. The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. *Statistics in Medicine.* 2013, **32**, 864–883.

[21] Lowe R, Barcellos C, Coelho C, Bailey T, Coelho G, Graham R, Jupp T, Massa-Ramalho W, Stephenson D, Rodó X. Dengue outlook for the World Cup in Brazil: an early warning model framework driven by real-time seasonal climate forecasts. *Lancet Infectious Diseases.* 2014, **14**, 619-626 .

[22] Lowe R, Cazelles B, Paul R, Rodó X. Quantifying the added value of climate information in a spatio-temporal dengue model. *Stochastic Environmental Research and Risk Assessment.* 2016, **30**, 2067–2078.

[23] Restrepo AC, Baker P, Clements ACA. National spatial and temporal patterns of notified dengue cases, Colombia 2007-2010. *Tropical Medicine and International Health.* 2014, **19**, 863–871.

[24] Arboleda S, Jaramillo-O, N, Peterson AT. Mapping Environmental Dimensions of Dengue Fever Transmission Risk in the Aburrá Valley, Colombia. *International Journal of Environmental Research and Public Health.* 2009, **6**, 3040–3055.

[25] Martínez-Bello DA, López-Quílez A, Torres-Prieto A Bayesian dynamic modeling of time series of dengue disease case counts.. *PLoS Neglected Tropical Diseases.* 2017, **11**, e0005696.

[26] Martínez-Bello DA, López-Quílez A and Torres Prieto A. Relative risk estimation of dengue disease at small spatial scale. *International Journal of Health Geographics.* 2017, **16**, 31.

[27] Martínez-Bello D, López-Quílez A, Torres Prieto A. Spatiotemporal modeling of relative risk of dengue disease in Colombia. *Stochastic Environmental Research and Risk Assessment.* 2018, **32**, 1587-1601.

[28] Stewart-Ibarra A, Muñoz Á, Ryan S, Ayala E, Borbor-Cordova M, Finkelstein J, Mejía R, Ordoñez T, Recalde-Coronel G, Rivero K. Spatiotemporal clustering climate periodicity and social-ecological risk factors for dengue during an outbreak in Machala, Ecuador, in 2010. *Infectious Diseases.* 2014, **14**, 610.

[29] Lippi CA, Stewart-Ibarra AM, Muñoz ÁG, Borbor-Cordova MJ, Mejía R, Rivero K, Castillo K, Cárdenas WB, Ryan SJ. The Social and Spatial Ecology of Dengue Presence and Burden during an Outbreak in Guayaquil, Ecuador, 2012. *International Journal of Environmental Research and Public Health.* 2018, **15**, 827.

[30] Wijayanti SPM, Porphyre T, Chase-Topping M, Rainey SM, McFarlane M, Schnettler E, Biek R, Kohl A. The Importance of Socio-Economic Versus Environmental Risk Factors for Reported Dengue Cases in Java, Indonesia. *PLoS Neglected Tropical Diseases.* 2016, **10**, 1-15.

[31] Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* 2009, **71**, 319–392.

[32] Knorr-Held L. Bayesian modelling of inseparable space-time variation in disease risk. *Statistics in Medicine.* 2000 **19**, 2555-2567.

[33] Abd Naeeim NS, Rahman NA. Estimating relative risk for dengue disease in Peninsular Malaysia using INLA. *Malaysian Journal of Fundamental and Applied Science.* 2017 **13**, 721-727.

[34] National Health Institute of Colombia. *Methodology of the routinary surveillance statistical operation.* Health Ministry of Colombia, National Health Institute, Bogotá. 2018. 92 pages. [In spanish]. URL: `https://www.ins.gov.co/Direcciones/Vigilancia/Lineamientosydocumentos/Metodolog%C3%ADa%20%20Sivigila.pdf`

[35] National Administrative Department of Statistics (DANE), Colombia. Direction of Geostatistics.*Urban Sector Level Layer, Department of Santander*. National Geostatistical Framework. 2005. [In spanish].URL: `https://geoportal.dane.gov.co/v2/index.php?page=elementoDescargaMGN`

[36] National Administrative Department of Statistics (DANE), Colombia. *Census 2005*. [In spanish]. URL: `http://systema59.dane.gov.co/cgibin/RpWebEngine.exe/PortalAction?&MODE=MAIN&BASE=CG2005BASICO&MAIN=WebServerMain.inl`

[37] Banerjee S, Carlin B, Gelfand A. *Hierarchical Modeling and Analyisis for Spatial Data*. Chapman & Hall/CRC biostatistics series: 6000 Broken Sound Parkway NW, Suite 300 Boca Raton, FL 33487-2742, 2015, p. 529.

[38] Leroux BG, Lei X, Breslow N. Estimation of disease rates in small areas: a new mixed model for spatial dependence. In: *Statistical Models in Epidemiology, the*

Environment and Clinical Trials; Halloran M, Berry D., Ed.; Springer-Verlag: New York, 1999; pp. 179–191.

[39] Ugarte MD, Adin A, Goicoa T, Militino AF. On fitting spatio-temporal disease mapping models using approximate Bayesian inference. *Statistical Methods in Medical Research.* 2014, **23**, 507-530.

[40] Ugarte M, Adin A, Goicoa T. Two-level spatially structured models in spatio-temporal disease mapping. *Statistical Methods in Medical Research.* 2016, **25**, 1080–1100.

[41] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. Url: https://www.R-project.org/

[42] Goicoa T, Adin A, Ugarte MD, Hodges JS. In spatio-temporal disease mapping models, identifiability constraints affect PQL and INLA results. *Stochastic Environmental Research and Risk Assessment.* 2018, **32**, 749-770.

[43] Watanabe S. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research.* 2010, **11**, 3571–3594

[44] Gneiting T, Raftery AE. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association.* 2007, **102**, 359–378

[45] Krystosik AR, Curtis A, Buritica P, Ajayakumar J, Squires R, Dávalos D, et al. Community context and sub-neighborhood scale detail to explain dengue, chikungunya and Zika patterns in Cali, Colombia. *PLoS ONE.* 2017, **12**, e0181208.

[46] Chien L-C, Lin R-T, Liao Y, Sy FS, Pérez, A. Surveillance on the endemic of Zika virus infection by meteorological factors in Colombia: a population-based spatial and temporal study. *BMC Infectious Diseases.* 2018, **18**, 180.

[47] Stewart-Ibarra, AM, Hargrave A, Diaz A, Kenneson A, Madden D, Romero MM, Molina JP, Macias Saltos D. Psychological Distress and Zika, Dengue and Chikungunya Symptoms Following the 2016 Earthquake in Bahía de Caráquez, Ecuador. *International Journal Environmental Research and Public Health.* 2017, **14**, 1516.

[48] Riou J, Chiara P, Boelle PY. A comparative analysis of Chikungunya and Zika transmission. *Epidemics.* 2017, **19**, 43-52.

[49] Perkins TA, Metcalf CJE., Grenfell BT, Tatem AJ. Estimating drivers of autochthonous transmission of Chikungunya virus in its invasion of the Americas. *PLoS Currents Outbreaks.* 2015, **7**. http://dx.doi.org/10.1371/currents.outbreaks.a4c7b6ac10e0420b1788c9767946d1fc

[50] Aguiar BS, Lorenz C, Virginio F, Suesdek L, Chiaravalloti-Neto F. Potential risks of Zika and chikungunya outbreaks in Brazil: A modeling study. *International Journal of Infectious Diseases.* 2018, **70**, 20-29.

[51] Lowe R, Barcellos C, Brasil P, Cruz OG, Honório NA, Kuper H, Carvalho MS. The Zika Virus Epidemic in Brazil: From Discovery to Future Implications. *International Journal of Environmental Research and Public Health.* 2018, **15**, 96.

[52] Saiz J-C, Martín-Acebes MA, Bueno-Marí R, Salomón OD, Villamil-Jiménez LC, Heukelbach J, Alencar CH, Armstrong PK, Ortiga-Carvalho TM, Mendez-Otero R, Rosado-de-Castro PH and Pimentel-Coelho PM. Zika Virus: What Have We Learnt Since the Start of the Recent Epidemic? *Frontiers in Microbiology.* 2017, **8**, 1554.

[53] Sallam MF, Fizer C, Pilant AN, Whung P-Y. Systematic Review: Land Cover, Meteorological, and Socioeconomic Determinants of Aedes Mosquito Habitat for Risk Mapping. *International Journal of Environmental Research and Public Health.* 2017, **14**, 1230.

[54] Romero-Vega L, Pacheco O, de la Hoz-Restrepo F, Díaz-Quijano FA. Evaluation of dengue fever reports during an epidemic, Colombia. *Revista de Sáude Pública.* 2014, **48**, 899–905.

[55] National Health Institute of Colombia. Subregister of Zika in Girardot, Cundinamarca, 2015-2016. [In spanish]. *Informe Quincenal Epidemiológico Nacional (IQEN).* 2016, **21** (21):501-522. `http://www.ins.gov.co/buscador-eventos/IQEN/IQEN%20vol%2021%202016%20num%2023.pdf`

Chapter 7

Joint estimation of relative risk of dengue and Zika Infections

Abstract

This study describes the joint estimation of relative risk of dengue and Zika virus disease (ZVD), establishing the spatial association between them at the departmental and the city level in Colombia, for the period of October 2015 to December 2016. Cases of dengue and ZVD were allocated to the 87 municipalities of one Colombian department and the 293 census sections of one Colombian city. Eight hierarchical Bayesian Poisson joint models of relative risk for dengue and ZVD were fitted to the data. The models included area- and disease-specific random effects accounting for several spatial patterns of disease risk (clustered or uncorrelated heterogeneity) within and between dengue and ZVD. The selected models captured the spatial uncorrelated patterns for dengue and ZVD high-risk and the shared spatial clustered patterns for both diseases at departmental level, while at the city level, spatial clustered patterns of dengue high-risk conditioned ZVD high-risk areas.

7.1 Introduction

Dengue and Zika virus disease (ZVD) are infectious arboviral diseases caused by Flavivirus, in the Flaviviridae family. Dengue virus has four serotypes (1, 2, 3, and 4). Serotypes 2 and 3 are associated with severe disease in second dengue infections, while Zika virus infection is associated with congenital malformations in babies born from women infected during pregnancy and Guillain-Barré syndrome in infected adults [1]. The country of Colombia located in South America is highly affected by vector-borne

diseases. Colombian health authorities reported over 101,016 dengue cases in 2016 resulting in 289 deaths ([2]), while there were 9,799 confirmed by laboratory - and 96,860 suspected – cases of ZVD by clinical signs to the end of 2016 [3]. In this study we concentrate on the spatial patterns assessment of risk of dengue and ZVD, and in particular we focus on the relative risk estimation for areal data using hierarchical Bayesian models for these events. Relative risk is a statistic representing the excess (or lack) of risk in a small area given a background risk. Relative risk is mostly model based and supported by Bayesian estimation methods [4]. We use as study regions, the municipalities belonging to the department of Santander in Colombia, which is one of the Colombian departments displaying the highest incidence of dengue and ZVD for the period 2015-2016, and the city census sections belonging to the capital city of the department of Santander, which is also one of the more affected cities by dengue and ZVD for the same period. The spatial patterns assessment of dengue risk has been reviewed by Racloz et al [5] and Louis et al [6], and for the specific case of relative risk estimation of dengue, Ferreira and Schmidt [7] and Martínez-Bello et al [8] estimated relative risk of dengue at local spatial scale, while Cadavid-Restrepo et al [9] and Martínez-Bello et al [10] applied methods for the spatio-temporal assessment of dengue risk. There are some examples for the spatial patterns assessment of ZVD risk, from the merely descriptive methods to the model based approaches. For instance, in Colombia, descriptive risk maps associating ZVD incidence rates with environmental and social factors have been produced in the departments of Sucre [11], Tolima [12], Guajira [13], Santander and Norte de Santander [14], and the city of Pereira in the department of Risaralda [15]. Model based approaches for the spatial pattern assessment of ZVD risk have been developed for the 33 departments in Colombia using Poisson models for the relative risk of ZVD including meteorological and environmental variables [16], while the distribution of risk to ZVD transmission among counties/districts in Guangdong Province, China was assessed using analytic hierarchy process (AHP) models [17]. ZVD, dengue and Chikungunya virus disease (CVD) are also jointly studied using spatial patterns assessment of risk because the viral agents share the same vectors (mosquitoes from the Aedes genus). For example, at large scale in Brazil, ecological studies have been developed to investigate the spatial risk factors of unusual patterns of microcephaly including dengue, ZVD and CVD data [18]; it has also been estimated the potential spatial risk of ZVD and CVD according to socio-environmental factors, and estimated the size of the populations at risk of both diseases [19]; and at small geographic scale, the risk factors for arboviral diseases (dengue, ZVD, CVD) at community level have been evaluated [20]. In the city of Cali, Colombia, Krystosik et al [21] applied spatial-video narratives in hotspots, generating risk maps of arboviral diseases (CVD, ZVD and dengue) at a local level in order to support vector-control strategies, while Martínez-Bello et al [22] estimated the relative risk of dengue and ZVD using spatio-temporal interaction effects model for one department and one city from Colombia, profiling a method for the joint estimation of dengue and ZVD high-risk. A modeling approach including spatial patterns assessment of risk using joint modeling of ZVD and CVD has been undertaken by Riou et al [23] analyzing the 2013 ZVD outbreaks in the French Polynesia islands, while the joint modeling of ZVD and dengue was developed by Funk et al [24] using time series data from the ZVD outbreak in the

Yap island in the Pacific ocean. The aim of this study is to jointly estimate the disease- and area-specific relative risk of dengue and ZVD using hierarchical Bayesian joint models accounting for the spatial association between both diseases. We use data for the 2015-2016 ZVD outbreak in Colombia, analyzing two levels of spatial data aggregation: the departmental level (disease counts aggregated per municipality) in the department of Santander, Colombia, and the city level (disease counts aggregated per census section) in the city of Bucaramanga (Santander).

7.2 Materials and methods

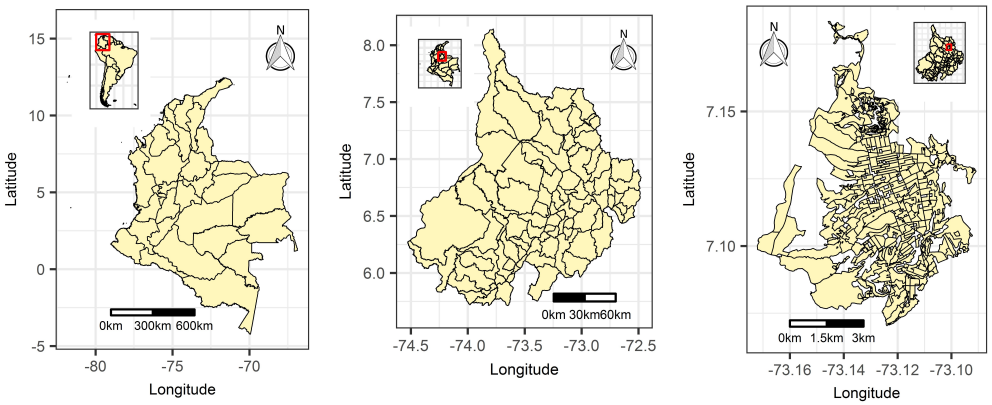
Colombia is located in the northwestern corner of South America (Figure 7.1, panel a). It has an area of about 1.14 million km2 and a population of 48.7 million, divided in 33 administrative regions called departments, which range in size from 49.6 km2 to 110,029.4 km2 (Table 7.1).

Table 7.1: Geographical division at national, departmental and municipal level.

		Area (km^2)			
Level	Division	Min.	Mean	Max.	Total
Colombia	33 departments	49.6	34,678.3	110,029.4	1,144,385
Santander	87 municipalities	18.8	352.2	3173.8	30,642
Bucaramanga	293 census sections (CS)	0.01	0.17	2.64	49.6

The department of Santander (Figure 7.1, panel b) covers an area of 30,642 km2 and is divided in 87 municipalities, with 2,071,016 inhabitants. The department is located in the northeast of Colombia, and Bucaramanga (Figure 7.1, panel c), its administrative center, is located in the northeastern region of the department. Its area of 162 km2 (urban area 49.6 km2) is divided in 293 urban census sections (CS), and according to the 2016 census, it has a population of 528,575.

The ZVD epidemic started on August 9, 2015 [25]. Up to the first epidemiological week (EW) of 2017, Colombia reported 106,659 ZVD cases (suspected by clinical signs plus confirmed by laboratory), or 219 cases per 100,000 population. For the epidemiological year 2016, the country reported 101,016 dengue cases (207.6 cases per 100,000 population; Table 2). We chose the department of Santander as study area because it presents the sixth incidence rate of dengue for year 2016 (331.63 cases per 100,000 population) [2], and the seventh incidence rate for ZVD for the period 201-2016 (493.09 cases per 100,000 population) [3] among the 33 Colombian departments. Santander registered a higher ZVD incidence rate than the national average, accounting for 9.76% of the total cases in Colombia [2]. With regard to dengue, Santander reported 6.67% of total cases for Colombia in 2016 (Table 2). Within the department, 35.52% of the ZVD cases



(a) Country of Colombia (b) Department of Santander (c) City of Bucaramanga

Figure 7.1: Geographic location of the study area.

(692.81 per 100,000 population) and 35.62% of the dengue cases (467.3 per 100,000 population) were reported in Bucaramanga (Table 7.2). Below we present the data sources at departmental and city level, before going on to describe the models of joint estimation of relative risk at both diseases, together with the inference and estimation methods.

7.2.1 Departmental level data

Cases of dengue and ZVD for the study period (from the 42nd EW of 2015 to the 52nd EW of 2016) obtained from the SIVIGILA (public health surveillance system) were aggregated in the 87 municipalities of Santander using the cartography from the National Geostatistical Framework [26]. The case definitions for dengue and ZVD correspond to the codes 210, 220 and 580 (dengue, severe dengue and dengue mortality respectively), and code 895 for ZVD, obtained from the SIVIGILA protocols for dengue [27] and ZVD [28]. Expected values of dengue and ZVD by municipality were calculated using 2016 population projections provided by the Colombian Administrative Department of National Statistics (DANE). We first calculated incidence rates of dengue and ZVD for five-year age bands and sex based on the observed cases recorded during the study period. We then multiplied these age- and sex-specific incidence rates to the population structure of each municipality, generating disease-specific expected values for both diseases. Finally, we aggregated the age- and sex-specific expected values per municipality of dengue and ZVD obtaining standardized expected counts per municipality for dengue and ZVD.

Table 7.2: Geographical division at national, departmental and municipal level.

Level	Suspected[a]	Confirmed[b]	Total[c]	Incidence rate[d]	Pop. x 1000
		ZVD			
Colombia[e]	96,860	9,799	106,659	219.0	48,654
Santander[f]	9,420	547	9,967	493.1	2,090
Bucaramanga[g]			3,662	692.8	529
		Dengue			
Colombia[g]	100,117	899	101,016	207.6	48,654
Santander[g]			6934	331.6	2,090
Bucaramanga[h]			2470	467.3	529

[a]Suspected cases by clinical signs.
[b]Laboratory-confirmed cases.
[c]Suspected cases with laboratory confirmation.
[d](Total cases/population x 1000) x 100,000.
[e]9 August 2015 to 5 January 2017, source: [3].
[f]EW 45 of 2015 to epidemiological week (EW) 52 of 2016, source: [3] .
[g]EW 1 to 52, 2016, source: [2].
[h]EW 45 of 2015 to EW 52 of 2016, source (SIVIGILA, authors).

7.2.2 Municipal level data

Cases of dengue and ZVD, notified in Bucaramanga and obtained from SIVIGILA for the study period, were geocoded and allocated to 293 CS in the city [26]. The CS is a cartographical unit comprised by 1 to 9 census blocks, consisting of built or unbuilt lots bounded by public roads or pedestrian walkways (26). We calculated expected values of dengue and ZVD by CS using a similar method to the one applied at the departmental level. We calculated incidence rates of dengue and ZVD by five-year age groups and sex for the study period, using the 2016 population projections provided by the Colombian statistical office. Next, we multiplied the incidence rates to the population structure (by sex and age group) of each CS obtained from the Colombian Census, and finally aggregating the age- and sex-specific expected counts per CS to obtain the standardized expected values of dengue and ZVD per CS.

7.2.3 Joint models for estimation of relative risk

Let us assume the observed counts of dengue or ZVD aggregated by area (87 municipalities at departmental level; or 293 CS for the city level) follow a Poisson distribution, with mean parameter equal to the product of the disease- and area- specific expected values and the relative risk. The logarithm of the relative risk is the additive result of disease- and area-specific random effects accounting by the uncorrelated and clustered spatial patterns of risk and possibly covariates. Random effects for spatial clustered patterns

of risk are unobserved variables recovering risk autocorrelation between adjacent areas, indicating that the risk in one area is highly associated with the risk in the neighboring areas. The lack of spatial risk autocorrelation is accounted for by spatial uncorrelated random effects [4] [29]. A joint model of relative risk for dengue and ZVD defines a structure for the nature of the spatial association between the diseases. To that end, we fitted eight Bayesian Poisson models [4] [30] of joint estimation of relative risk to the dengue and ZVD data at departmental and city level. We kindly suggest to the readers to check the statistical model formulation available as a supplementary document posted online (https://github.com/danieladyro/SupplementaryMaterialEID). This section is limited to presenting and describing the spatial patterns of risk of every joint model:

- Model 1 contains area- and disease-specific random effects capturing spatial uncorrelated patterns of dengue and ZVD risk. Choosing this model implies that dengue and ZVD high-risk areas are not associated with each other and show no clustering (Figure 7.2, panel a).

- Model 2 incorporates area- and disease-specific random effects capturing spatial uncorrelated patterns of dengue and ZVD risk, where the random effects are spatially associated. Model 2 assumes that the random effects for spatial uncorrelated patterns are linearly associated for both diseases, so dengue and ZVD high-risk areas occur at the same locations (Figure 7.2, panel b).

- Model 3 contains area- and disease-specific random effects representing spatial clustered patterns of dengue and ZVD risk, where the random effects are not associated between diseases. This models reveals that dengue (ZVD) high-risk areas are strongly associated with other dengue (ZVD) high-risk areas (Figure 7.2, panel c).

- Model 4 contains area- and disease-specific random effects revealing spatial clustered patterns of dengue and ZVD risk, where the random effects are linearly associated between diseases. In this model dengue and ZVD high-risk areas are spatially associated, so there are spatial clustered patterns of dengue and ZVD high-risk areas at the same spatial locations (Figure 7.2, panel d).

- Model 5 incorporates area- and disease-specific random effects revealing spatial uncorrelated patterns of dengue and ZVD risk, and the spatial association between both diseases is modeled using shared-components of clustered spatial patterns of dengue and ZVD risk. This model reveals that dengue high-risk areas are neighboring ZVD high risk areas, but dengue (ZVD) high-risk areas are not adjacent to other dengue (ZVD) high-risk areas (Figure 7.2, panel 3).

- Model 6 incorporates area- and disease-specific random effects revealing spatial clustered patterns of dengue and ZVD risk, and the spatial association between both diseases is modeled using shared random effects of spatial clustered patterns of dengue and ZVD risk. Selecting this model reveals that spatial clustered patterns

of dengue are adjacent to spatial clustered patterns of ZVD risk (Figure 7.2, panel f).

- Model 7 contains area- and disease-specific random effects accounting for spatial clustered patterns of ZVD risk conditioned on random effects for spatial clustered patterns of dengue risk, so dengue high-risk areas are determinants of the presence of ZVD high-risk areas (Figure 7.2, panel g).

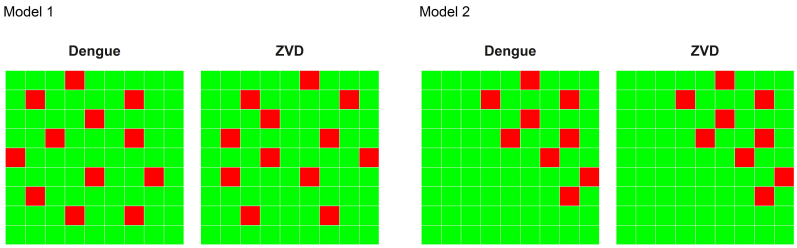
- Model 8 contains area- and disease-specific random effects accounting for spatial clustered patterns of dengue risk conditioned on random effects for spatial clustered patterns of ZVD risk, so ZVD high-risk areas are determinants of the presence of dengue high-risk areas (Figure 7.2, panel h).

Table 7.3 summarizes the association structure of the joint models of relative risk, so every model captures the nature of the spatial association between dengue and ZVD risk. The Bayesian Poisson models were fitted applying Markov chain Monte Carlo simulations using WinBUGS 1.4 [31] software for Bayesian analysis. Non-informative prior distributions were selected for the parameters and hyper-parameters; model specifications are presented in the supplementary material. Every model was fitted using 100,000 burn-in iterations, 10,000 iterations for inferences, a thinning of 10, three chains, and 1000 iterations per chain for final inferences. We employed deviance information criterion (DIC) [32] and the effective number of parameters and the mean deviance. The final selected model corresponds to the model displaying the lowest DIC.

Table 7.3: Association structure assumed by the relative risk models 1 to 8 fitted to the departmental and municipal level dengue and ZVD data for the 2015-2016 outbreak.

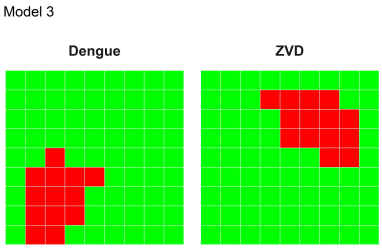
Model	Spatially structured association of dengue and ZVD* high-risk areas	Joint association between dengue and ZVD high-risk areas
1	No	No
2	No	Yes, linear
3	Yes	No
4	Yes	Yes, linear
5	No	Spatially structured shared component
6	Yes	Spatially structured shared component
7	Yes	ZVD risk conditioned by dengue risk
8	Yes	Dengue risk conditioned by ZVD risk

*ZVD: Zika virus disease

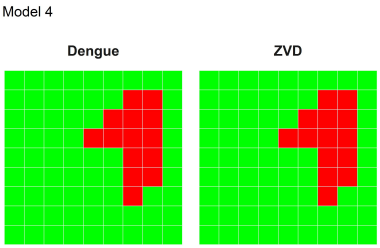


(a) Model 1

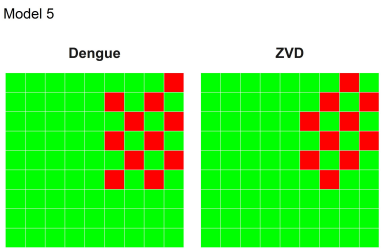
(b) Model 2



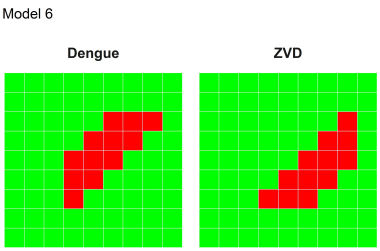
(c) Model 3



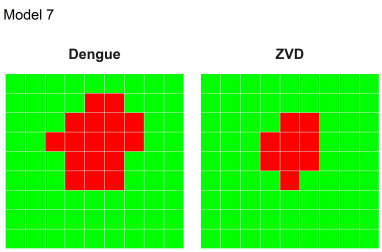
(d) Model 4



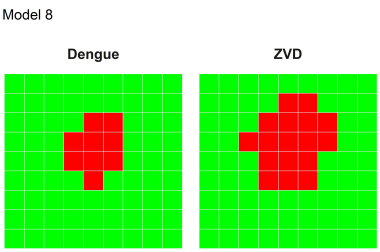
(e) Model 5



(f) Model 6



(g) Model 7

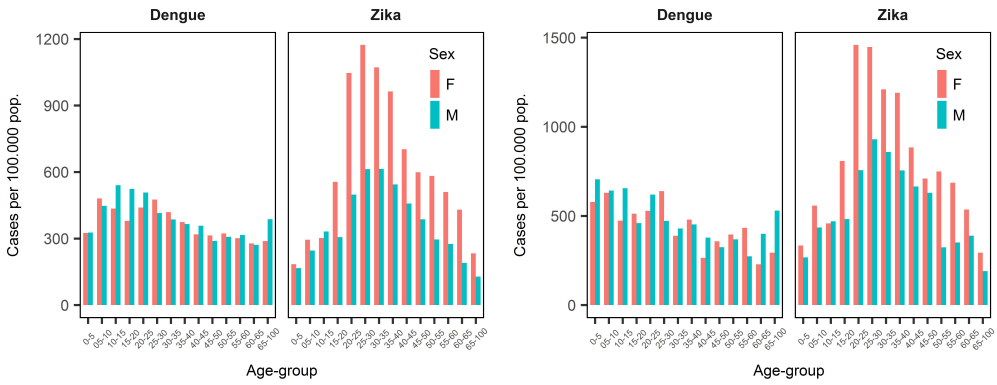


(h) Model 8

Figure 7.2: Schematic representation of the spatial patterns of dengue and ZVD risk captured by the joint models of relative risk.

7.3 Results

For the department of Santander, the analysis includes 10,051 ZVD cases (63.1% females and 36.9% males) and 7891 dengue cases (48.6% females and 51.4% males). The analysis of the city of Bucaramanga includes 3,662 ZVD cases (61.2% females and 38.8% males) and 2,470 dengue cases (49.3% females and 50.7% males). Figure 7.3 shows the age- and sex-specific incidence rates for dengue and ZVD in Santander (Figure 7.3, panel a) and Bucaramanga (Figure 7.3, panel b). At departmental and city levels, females in the 10–14 and 55–59 age groups consistently reported more ZVD cases than males, while the reported dengue cases were the highest in the 5–9 and 20–24 age groups for both sexes. The age- and sex-specific incidence rates were slightly higher at city level than departmental level. Figure 7.4 maps the incidence rates (IR) (cases per 100,000 population) and the standardized incidence ratios (SIR) (observed values/expected values) of dengue and ZVD. At departmental level, the IR for ZVD ranges from 0 to 3688 per 100,000 population, and for dengue from 0 to 4285 cases per 100,000 population (Figure 7.4a). At departmental level, the highest IR for dengue was in the southeastern municipalities, while the highest IR for ZVD was in the northeastern municipalities. The SIR for dengue and ZVD at departmental level follows the findings of the IR map; however, the SIR accounts for the expected number of cases, so some areas with a high observed IR do not show a high SIR (Figure 7.4, panel b). Figure 7.4, panel c shows heterogeneous IRs of dengue and ZVD at city level, with some high IR CS in the northern and central areas of the city. Figure 7.4, panel d displays the SIR per CS, and as before, the SIR map recovers the findings of the IR, smoothing the incidences by CS in Bucaramanga. Table 7.4 presents the results of the selection statistics for the joint models

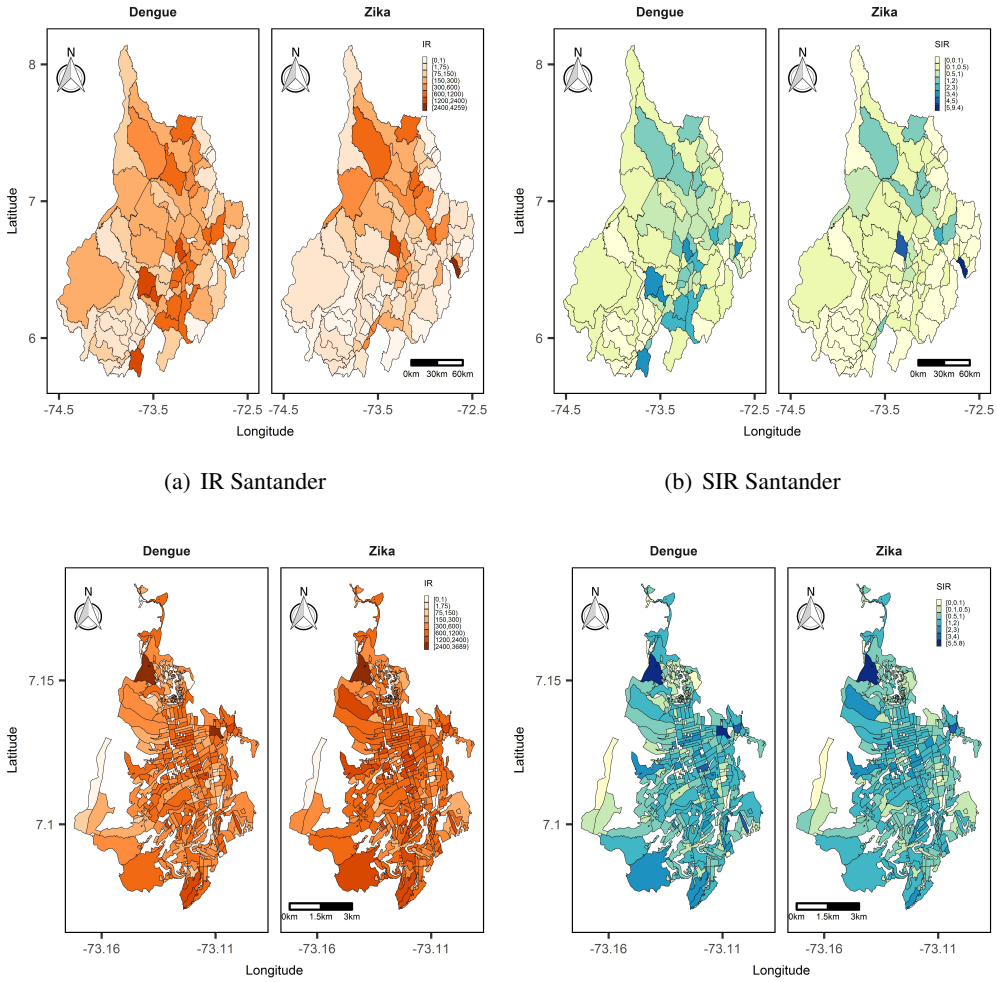


(a) Department of Santander

(b) City of Bucaramanga

Figure 7.3: Incidence rate of dengue and Zika virus disease per 100,000 population by age-group and sex, 2015-2016.

of relative risk at departmental level. Based on the lowest DIC, Model 5 (disease-specific random effects for uncorrelated spatial patterns of risk, and shared random effects of



(a) IR Santander

(b) SIR Santander

(c) IR Bucaramanga

(d) SIR Bucaramanga

Figure 7.4: Maps of the incidence rate (IR) and standardized incidence rate (SIR), 2015-2016.

spatial clustered patterns for both diseases) is the selected model (deviance = 808.3, DIC = 942.9) for Santander while, at city level (Table 7.5), model 7 (ZVD risk distribution conditioned by dengue risk) is the selected final model (deviance = 2869.7, DIC = 3119.3) for Bucaramanga. In Table 7.4 and Table 7.5, we include the DIC differences between models (Δ-DIC), showing that after the final selected model, models 1 and 8 were the closest models at departmental level, while models 2 and 4 at city level. Figure 7.5 displays the posterior mean relative risk (RR) and the lower bound of the 95% credible interval (CI) of the RR greater than 1 (95% CI RR $>$ 1) of dengue and ZVD obtained from the selected final models in Table 7.4 and Table 7.5. In Santander, the posterior mean RR in Figure 7.5, panel a recovers the SIR pattern displayed in Figure 7.4, panel a,

Table 7.4: Deviance, effective number of parameters, DIC, and Δ DIC from the joint models of dengue and Zika for the department of Santander

				Δ − DIC						
Model	$\bar{D}$	p_D	DIC	M2	M3	M4	M5	M6	M7	M8
1	810.9	144.6	955.5	1.3	359.9	321.6	12.6	330.7	5.5	4.1
2	814.2	140.0	954.2		361.2	322.9	11.3	332	4.2	2.8
3	991.4	324.0	1315.4			38.3	372.5	29.2	365.4	364
4	976.8	300.4	1277.1				334.2	9.1	327.1	325.7
5	808.3	134.6	942.9					343.3	7.1	8.5
6	987.3	299.0	1286.2						336.2	334.8
7	810.4	139.6	950.0							1.4
8	811.7	139.7	951.4							

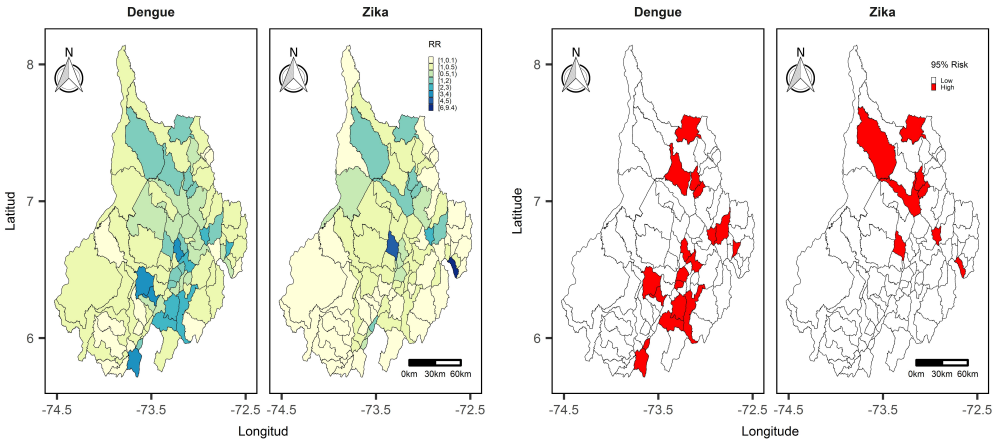
Table 7.5: Deviance ($\bar{D}$), effective number of parameters (p_D), deviance information criteria (DIC), and Δ DIC from the joint models of dengue and Zika for the city of Bucaramanga

				Δ − DIC						
Model	$\bar{D}$	p_D	DIC	M2	M3	M4	M5	M6	M7	M8
1	2870.1	364.9	3232.1	99.8	22.3	101.3	97.9	100	112.8	102.9
2	2853.3	279.0	3132.3		122.1	1.5	1.9	0.2	13	3.1
3	2922.4	331.9	3254.4			123.6	120.2	122.3	135.1	125.2
4	2884.2	246.6	3130.8				3.4	1.3	11.5	1.6
5	2856.8	277.4	3134.2					2.1	14.9	5
6	2888.9	243.3	3132.1						12.8	2.9
7	2869.7	249.6	3119.3							9.9
8	2874.2	255.0	3129.2							

and as a by-product of the modeling process, Figure 7.5, panel b shows the municipalities with 95% probability of RR higher than the other municipalities. Most of the dengue and ZVD high-risk municipalities differ in their risk distribution: while ZVD high-risk municipalities are located in the north of the department, dengue high-risk municipalities are in the south to northeast.

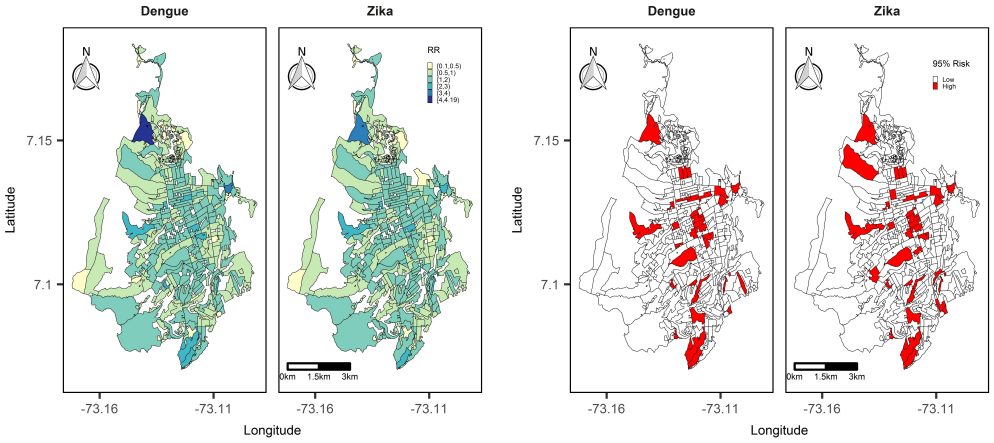
In Bucaramanga Figure 7.5, panel c shows the dengue and ZVD posterior mean RR maps recovering the non-clustered risk pattern of the diseases, also displayed by the SIR map in Figure 7.4, panel c. However, the model shrinks the posterior mean RR, capturing the close association between ZVD and dengue high-risk distribution per CS.

Figure 7.5, panel d presents the CSs having 95% probability of RR higher than the other areas, showing dengue high-risk CSs associated with ZVD high-risk CSs at city level.



(a) Posterior mean RR of Santander

(b) 95 % CI RR $>$ 1 of Santander



(c) Posterior mean RR of Bucaramanga

(d) 95% CI RR $>$ 1 of Bucaramanga

Figure 7.5: Posterior mean relative risk (RR) and 95% credible interval (CI) of RR greater than 1 (95% CI RR $>$ 1), from model 5 for the department of Santander and from model 7 for the city of Bucaramanga, 2015-2016.

7.4 Discussion

The present paper illustrates the joint estimation of relative risk of dengue and ZVD at departmental and city level in Colombia. A battery of joint models of relative risk were fitted to the ZVD and dengue data capturing the spatial association between the diseases during the 2015-2016 ZVD outbreak. The model selection process was based on the DIC statistics, and the model's goodness of fit were assessed using residual analysis, fitted-observed scatter plots and the posterior predictive checks of over-dispersion recovery, as it is shown in the supplementary material.

At departmental level, the final model contains random effects accounting for spatial clustered patterns of risk for both diseases considered as one entity and disease-specific spatial uncorrelated random effects for dengue and ZVD. Thus, the dengue- and ZVD-specific risk per municipality present spatial uncorrelated patterns, with dengue high risk municipalities being neighbors of ZVD high risk municipalities, but without dengue (ZVD) high risk municipalities neighboring other dengue (ZVD) high-risk municipalities.

At city level, the final model contains conditional random effects for spatial clustered patterns of risk [4], where ZVD high-risk CSs are conditioned on dengue high-risk CSs, meaning that ZVD risk by CS is highly associated and conditioned by dengue risk by CS. The selected model establishes that the dengue or ZVD risk present spatial clustered patterns, so dengue and ZVD high-risk CSs are near other high-risk CSs. The selected model implies that within the city, there are CS clusters displaying favorable conditions for the transmission of both diseases, justifying a posterior analysis of the environmental, infrastructural and socioeconomic conditions associated with dengue and ZVD high risk in those areas.

The risk maps generated provide model-based estimates of posterior means and credible intervals of relative risk, identifying areas with a given risk probability and highlighting municipalities or CSs at high probable risk for dengue or ZVD. The joint models surpass the disease-specific risk estimates presented at small scale by Krystosik et al. [21], establishing an objective measure of spatial association between the diseases. Our study also extends joint modelling of times series of dengue and ZVD such as Funk et al's [24] study, using the spatial nature of the data.

The present study has some limitations. Although dengue and ZVD are compulsory notifiable diseases, a large proportion of reported cases are diagnosed by clinical signs but not confirmed by laboratory. Our data correspond to both the suspected and confirmed dengue and ZVD cases, meaning that some of the cases included in the study may have been misdiagnoses. Underreporting is another potential source of bias in the study. Romero-Vega et al [33] estimated underreporting for dengue in Colombia, while the National Institute of Health of Colombia calculated an underreporting rate of 49% for ZVD in a high-incidence Colombian town [34]. Even with the current limitations, we need to remark that we are using the same data that the Colombian health authorities use to provide health information for decision making of controlling and preventing activities of public and private health institutions, so with the models shown in the study, we extend the current information generated by the public health surveillance

system, supporting surveillance activities for ZVD and dengue, providing information for monitoring the geographical distribution of the diseases and for spatially characterizing the disease distribution in the population [35]. For future studies, we certainly need to include determinants of dengue and ZVD risk as covariates within the joint models (for instance using datasets such as the ecological and environmental spatial dataset developed by Sijraj et al [36] for Colombia), testing other joint models available in the spatial analysis literature, jointly modeling other arboviral diseases such as CVD together with dengue and ZVD, and including joint models of RR of dengue and ZVD within real-time surveillance platforms such as the one described by Jaramillo-Martínez et al [37] in a dengue an ZVD high-incidence city in Colombia.

7.5 References

[1] Plourde AR, Bloch EM. A Literature Review of Zika Virus. *Emerging Infect Dis*. 2016; **22**(7):1185-1192.

[2] National Health Institute. *Event report: dengue, 2016.* National Health Institute, Public Health Surveillance and Risk Analysis. [In spanish.] Bogotá D.C., Colombia. URL: `https://www.ins.gov.co/buscador-eventos/Informesdeevento/Dengue%202016.pdf`

[3] National Health Institute. Weekly epidemiological bulletin number 52 of 2016, 25 December – 31 December. National Health Institute, Public Health Surveillance and Risk Analysis Office, Bogotá D.C, Colombia. In spanish. 2016. Url: `www.ins.gov.co`

[4] Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data, Second Edition* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton, Fl.; 2014. 584 Pages.

[5] Racloz V, Ramsey R, Tong S, Hu W. Surveillance of dengue fever virus: A review of epidemiological models and early warning systems. *PLoS Negl Trop Dis.* 2012; **6**(5): 1648.

[6] Louis VR, Phalkey R, Horstick O, Ratanawong P, Wilder-Smith A, Tozan Y, et al. Modeling tools for dengue risk mapping - a systematic review. *Int J Health Geogr.* 2014; **13**(1): 50.

[7] Ferreira GS and Schmidt AM. Spatial modelling of the relative risk of dengue fever in Rio de Janeiro for the epidemic period between 2001 and 2002. *Braz J Probab Stat.* 2006; **20**: 29-47.

[8] Martínez-Bello DA, López-Quílez A and Torres Prieto. A Relative risk estimation of dengue disease at small spatial scale. *Int J Health Geogr.* 2017; **16**:31.

[9] Cadavid-Restrepo, A, Baker P, Clements ACA. National spatial and temporal patterns of notified dengue cases, Colombia 2007-2010. *Trop Med & Int Health.* 2014; **19**(7) :863-871

[10] Martínez-Bello, D.; López-Quílez, A.; Torres Prieto, A. Spatiotemporal modeling of relative risk of dengue disease in Colombia. *Stoch Environ Res Risk Assess.* 2018, **32**: 1587-1601

[11] Rodriguez-Morales AJ, Patiño-Cadavid LJ, Lozada-Riasco CO, Villamil-Gómez WE. Mapping Zika in municipalities of one coastal department of Colombia (Sucre) using geographic information systems during the 2015-2016 outbreak: implications for public health and travel advice. *Int J Infect Dis.* 2016; **48**: 70-72.

[12] Rodriguez-Morales AJ, Galindo-Marquez ML, García-Loaiza CJ et al. Mapping Zika virus infection using geographical information systems in Tolima, Colombia, 2015-2016 [version 1; referees: 2 approved] *F1000Research.* 2016; **5**:568.

[13] Rodriguez-Morales AJ, García-Loaiza CJ, Galindo-Marquez ML Sabogal-Roman JA, Marin-Loaiza S, Lozada-Riascos CO, Díaz-Quijano FA. Zika infection GIS-based mapping suggest high transmission activity in the border area of La Guajira, Colombia, a northeastern coast Caribbean department, 2015-2016: Implications for public health, migration and travel. *Travel Med Infect Dis.* 2016; **14**: 286-288.

[14] Rodriguez-Morales AJ, Haque U, Ball JD, García-Loaiza CJ, Galindo-Marquez ML, Sabogal-Roman JA, Marin-Loaiza S, Ayala AF, Lozada-Riascos CO, Diaz-Quijano FA, Alvarado-Socarras JL. Spatial distribution of Zika virus infection in Northeastern Colombia. *Infez Med.* 2017; **3**: 241-246.

[15] Rodriguez-Morales AJ, Ruiz P, Tabares J, Ossa CA, Yepes-Echeverry MC, Ramirez-Jaramillo V, Galindo-Marquez ML, García-Loaiza CJ, Sabogal-Roman JA, Parra-Valencia E, Lagos-Grisales GJ, Lozada-Riascos CO, de Pijper CA, Grobusch M. Mapping the ecoepidemiology of Zika virus infection in urban and rural areas of Pereira, Risaralda, Colombia, 2015-2016: Implications for public health and travel medicine. *Travel Med Infect Dis.* 2017; **18**: 57-66.

[16] Chien L-C, Lin R-T, Liao Y, Sy FS and Pérez A. Surveillance on the endemic of Zika virus infection by meteorological factors in Colombia: a population-based spatial and temporal study. *BMC Infect Dis.* 2018; **18**:180.

[17] Li X, Liu T, Lin L, Song T, Du X, Lin H, et al. Application of the analytic hierarchy approach to the risk assessment of Zika virus disease transmission in Guangdong Province, China. *BMC Infect Dis.* 2017; **17**:65 DOI 10.1186/s12879-016-2170-2

[18] Campos MC, Dombrowski JG, Phelan J, Marinho CRF, Hibberd M, Clark TG, et al. Zika might not be acting alone: Using an ecological study approach to investigate potential co-acting risk factors for an unusual pattern of microcephaly in Brazil. *PLoS ONE.* 2018; **13**(8): e0201452.

[19] Aguiar B, Lorenz C, Virginio F, Suesdek L, Chiaravalloti-Neto F. Potential risks of Zika and chikungunya outbreaks in Brazil: A modeling study. *Int J Infect Dis.* 2018; **70**: 20-9.

[20] Rodrigues NCP, Daumas RP, de Almeida AS, dos Santos RS, Koster I, Rodrigues PP, et al. Risk factors for arbovirus infections in a low-income community of Rio de Janeiro, Brazil, 2015-2016. *PLoS ONE.* 2018; **13**(6): e0198357.

[21] Krystosik AR, Curtis A, Buritica P, Ajayakumar J, Squires R, Dávalos D, et al. Community context and sub-neighborhood scale detail to explain dengue, chikungunya and Zika patterns in Cali, Colombia. *PLoS ONE.* 2017; **12**(8): e0181208.

[22] Martínez-Bello DA, López-Quílez A, Torres Prieto A. Spatio-Temporal Modeling of Zika and Dengue Infections within Colombia. *Int J Environ Res Public Health.* 2018; **15**: 1376.

[23] Riou J, Poletto C, Boëlle P. A comparative analysis of Chikungunya and Zika transmission. *Epidemics*. 2017; **19**: 43-52.

[24] Funk S, Kucharski AJ, Camacho A, Eggo RM, Yakob L, Murray LM, et al. Comparative Analysis of Dengue and Zika Outbreaks Reveals Differences by Setting and Virus. *PLoS Negl Trop Dis.* 2016 **10**(12): e0005173.

[25] Pacheco O, Beltrán M, Nelson CA, et al. Zika Virus Disease in Colombia - Preliminary Report. *N Engl J Med.* 2016. doi:10.1056/NEJMoa1604037.

[26] Departamento Administrativo Nacional de Estadística, Colombia. Dirección de Geoestadística. [National Administrative Department of Statistics, Colombia. Direction of Geostatistics]. Capa del Nivel de Sección Urbano [Urban Section Level Layer]. Marco Geoestadístico Nacional [National Geostatistical Framework]; 2005. Url: `http://www.dane.gov.co/`

[27] National Health Institute of Colombia. Protocol of public health surveillance, Dengue, code: 210, 220, 580. Health Ministry of Colombia, National Health Institute, Bogotá. 2017. 19 pages. [In spanish.] `https://www.ins.gov.co/buscador-eventos/ZIKA%20Lineamientos/Dengue%20PROTOCOLO.pdf`

[28] National Health Institute of Colombia. Protocol of public health surveillance, Zika virus disease, code 895. Health Ministry of Colombia, National Health Institute, Bogotáa. 2017. 22 pages. [In spanish]. `https://www.ins.gov.co/buscador-eventos/ZIKA%20Lineamientos/Zika%20PROTOCOLO.pdf`

[29] Besag J, York J, Mollie A. Bayesian image restoration with two applications in spatial statistics. *Annals Inst Stat Math.* 1991; **43**(1):1-59.

[30] Ma H and Carlin BP (2005) Bayesian Multivariate Areal Wombling for Multiple Disease Boundary Analysis. Technical Report. School of Public Health, University of Minnesota. url: `http://www.biostat.umn.edu/~brad/software/mc.pdf`

[31] Lunn D, Spiegelhalter D, Thomas A, Best N. The BUGS project: Evolution, critique, and future directions. *Stat Med.* 2009; **28**: 3049-3067.

[32] Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. *J Royal Stat Soc, Series B (Statistical Methodology).* 2002; **64**: 583-639.

[33] Romero-Vega L, Pacheco O, de la Hoz-Restrepo F, Díaz-Quijano FA. Evaluation of dengue fever reports during an epidemic, Colombia. *Rev Saude Pública.* 2014; **48**(6): 899-905.

[34] National Health Institute of Colombia. Subregister of Zika in Girardot, Cundinamarca, 2015-2016. Informe Quincenal Epidemiológico Nacional (IQEN), 2016; **21**(23): 501-522. In spanish. Url:`http://www.ins.gov.co/buscador-eventos/IQEN/Forms/AllItems.aspx`

[35] Saiz J-C, Martín-Acebes MA, Bueno-Marí R, Salomón OD, Villamil-Jiménez LC, Heukelbach J, et al. Zika Virus: What Have We Learnt Since the Start of the Recent Epidemic? *Front. Microbiol.* 2017; **8**:1554.

[36] Siraj AS, Rodriguez-Barraquer I, Barker CM, Tejedor-Garavito N, Harding D, Lorton C, et al. Data Descriptor: Spatiotemporal incidence of Zika and associated environmental drivers for the 2015-2016 epidemic in Colombia. *Scientific Data.* 2018; **5**: 180073.

[37] Jaramillo-Martínez GA, Vasquez-Serna H, Chavarro-Ordoñez R, Rojas-Gomez OF, Jimenez-Canizales CE, Rodriguez-Morales AJ. Ibagué Saludable: A novel tool of Information and Communication Technologies for surveillance, prevention and control of dengue, chikungunya, Zika and other vector-borne diseases in Colombia. *J Infect Public Health.* 2018; **11**: 145-146.

7.6 Supplementary material

Introduction

In this supplementary material we present to the interested reader the statistical formulation for the joint models in the paper "Joint estimation of relative risk of dengue and Zika infections" and the model diagnostic measures. Together with the supplementary material, the reader can use the .txt files located at the same repository of this file, containing the WinBUGS 1.4 code, data and initial values necessary to reproduce the results shown in the manuscript as a result of fitting the joint models of relative risk 1 to 8 for the department of Santander, and the city of Bucaramanga.

Statistical formulation of the joint models of relative risk

Let us assume the observed counts O_{ij} of dengue or Zika virus disease (ZVD) are Poisson distributed with mean parameter (μ_{ij}) where i is the aggregation area ($i = 1, \ldots, n$, and $n = 87$ municipalities at departmental level; or $n = 293$ census section for the municipal level), and j is the disease ($j = 1, 2$, and $p = 1$ for ZVD, or $p = 2$ for dengue).

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\mu_{ij} &= E_{ij} \times \text{r}_{ij} \\
\text{r}_{ij} &= \exp(\lambda_{ij})
\end{aligned}
$$

Then, the mean parameter μ_{ij} is equal to the product of the expected values E_{ij} and the relative risk r_{ij}, with linear predictor λ_{ij}. The logarithm of the relative risk is the additive result of spatially structured and unstructured random effects and covariates. Spatially structured random effects [1] are unobserved variables recovering a clustered risk pattern, or the fact that the risk in one area is highly associated with the neighboring areas. The lack of spatial association is accounted by the spatially unstructured random effects.

[1] Banerjee S, Carlin BP, Gelfand AE. *Hierarchical Modeling and Analysis for Spatial Data, Second Edition* Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton, Fl.; 2014. 584 Pages.

Model 1

Model 1 contains independent and identically distributed (IID) Normal spatially unstructured random effects for every disease (dengue and ZVD)[2].

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{ij}) &= \log(E_{ij}) + \alpha_j + \phi_{ij} \\
\boldsymbol{\phi}_j &\sim \text{Normal}(\mathbf{0}, \sigma^2_{\phi_j}\mathbf{I}) \\
\alpha_j &\sim \text{Normal}(0, 1000) \\
1/\sigma^2_{\phi_j} &\sim \text{Gamma}(0.01, 0.01)
\end{aligned}
$$

where $\boldsymbol{\phi}_j$ are spatially unstructured random effects, α_j are intercepts, $\sigma^2_{\phi_j}$ are variance parameters of the $\boldsymbol{\phi}_j$, and $\mathbf{I}$ is $n \times n$ identity matrix.

Model 2

Model 2 presents IID Normal spatially unstructured random effects linearly correlated for both diseases.

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{ij}) &= \log(E_{ij}) + \alpha_j + \phi_{ij} \\
\begin{bmatrix}\boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2\end{bmatrix} &\sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma} \otimes \mathbf{I}) \\
\boldsymbol{\Sigma} &= \begin{bmatrix} \sigma^2_1 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma^2_2 \end{bmatrix} \\
\boldsymbol{\Sigma}^{-1} &\sim \text{Wishart}(\mathbf{R}_{\mathbf{2}\times\mathbf{2}}, 2) \\
\boldsymbol{R} &= \begin{bmatrix} 1/5 & 0 \\ 0 & 1/5 \end{bmatrix} \\
\alpha_j &\sim \text{Normal}(0, 1000)
\end{aligned}
$$

$\boldsymbol{\Sigma}$ is an $n \times n$ variance-covariance matrix accounting for the association of the spatially unstructured random effects $\boldsymbol{\phi}_j$, $\otimes$ corresponds to the Kronecker product, and the rest of the parameters similar to model 1.

[2]Ma H and Carlin BP. Bayesian Multivariate Areal Wombling for Multiple Disease Boundary Analysis. Technical Report. School of Public Health, University of Minnesota. 2005. url: http://www.biostat.umn.edu/ brad/software/mc.pdf

Model 3

Model 3 accommodates conditionally autoregressive (CAR) Normal spatially structured random effects for every disease.

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{ij}) &= \log(E_{ij}) + \alpha_j + \phi_{ij} \\
\boldsymbol{\phi}_j &\sim \text{Normal}(\mathbf{0}, \sigma^2_{\phi_j}(\boldsymbol{D} - \boldsymbol{W})^{-1}) \\
\alpha_j &\sim \text{Normal}(0, 1000) \\
1/\sigma^2_{\phi_j} &\sim \text{Gamma}(0.01, 0.01) \\
\boldsymbol{W}_{n\times n} &= \begin{cases} w_{ij} = 1 & \text{if } i \sim j \\ w_{ij} = 0 & \text{if } i = j \\ w_{ij} = 0 & \text{otherwise} \end{cases} \\
w_{i+} &= \sum_j w_{ij} \\
\boldsymbol{D}_{n\times n} &= \text{diagonal}(w_{1+}, \ldots, w_{n+})
\end{aligned}
$$

where $\boldsymbol{\phi}_j$ are spatially structured random effects, with structure given by the proximity matrix $\boldsymbol{W}$, and variance parameters $\sigma^2_{\phi_j}$.

Model 4

Model 4 contains CAR Normal spatially structured random effects linearly correlated for both diseases.

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{ij}) &= \log(E_{ij}) + \alpha_j + \phi_{ij} \\
\begin{bmatrix} \boldsymbol{\phi}_1 \\ \boldsymbol{\phi}_2 \end{bmatrix} &\sim \text{Normal}(\mathbf{0}, \boldsymbol{\Sigma} \otimes (\boldsymbol{D} - \boldsymbol{W})^{-1}) \\
\boldsymbol{\Sigma} &= \begin{bmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix} \\
\boldsymbol{\Sigma}^{-1} &\sim \text{Wishart}(\mathbf{R}_{\mathbf{2}\times\mathbf{2}}, 2) \\
\boldsymbol{R} &= \begin{bmatrix} 1/5 & 0 \\ 0 & 1/5 \end{bmatrix} \\
\alpha_j &\sim \text{Normal}(0, 1000) \\
\boldsymbol{W}_{n\times n} &= \begin{cases} w_{ij} = 1 & \text{if } i \sim j \\ w_{ij} = 0 & \text{if } i = j \\ w_{ij} = 0 & \text{otherwise} \end{cases} \\
w_{i+} &= \sum_j w_{ij} \\
\boldsymbol{D}_{n\times n} &= \text{diagonal}(w_{1+}, \ldots, w_{n+})
\end{aligned}
$$

where $\boldsymbol{\phi}_j$ are spatially structured random effects, $\boldsymbol{\Sigma}$ is the variance-covariance matrix accounting the spatial association of the $\phi_{\mathbf{j}}$, and proximity matrix $\boldsymbol{W}$.

Model 5

Model 5 includes IID Normal spatially unstructured random effects for every disease with a CAR shared-parameter.

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{i,1}) &= \log(E_{i,1}) + \alpha_1 + \psi_i \times \gamma + \phi_{i,1} \\
\log(\mu_{i,2}) &= \log(E_{i,2}) + \alpha_2 + \psi_i / \gamma + \phi_{i,2} \\
\boldsymbol{\psi} &\sim \text{Normal}(\mathbf{0}, \sigma_\psi^2 (\boldsymbol{D} - \boldsymbol{W})^{-1}) \\
\boldsymbol{\phi}_j &\sim \text{Normal}(\mathbf{0}, \sigma_{\phi_j}^2 \mathbf{I}) \\
\alpha_j &\sim \text{Normal}(0, 1000) \\
\sigma_\psi &\sim \text{Uniform}(0, 10) \\
\sigma_{\phi_j} &\sim \text{Uniform}(0, 10) \\
\gamma &\sim \text{Normal}(0, 100) \\
\boldsymbol{W}_{n\times n} &= \begin{cases} w_{ij} = 1 & \text{if } i \sim j \\ w_{ij} = 0 & \text{if } i = j \\ w_{ij} = 0 & \text{otherwise} \end{cases} \\
w_{i+} &= \sum_j w_{ij} \\
\boldsymbol{D}_{n\times n} &= \text{diagonal}(w_{1+}, \ldots, w_{n+})
\end{aligned}
$$

$\boldsymbol{\psi}$ are the spatially structured shared-parameter components, γ is a scaling parameter, $\boldsymbol{\phi}_j$ are spatially unstructured random effects, $\boldsymbol{W}$ is the proximity matrix, and $\sigma_{\phi_j}^2$, σ_ψ^2 are variance parameters for $\boldsymbol{\phi}_j$ and $\boldsymbol{\psi}$.

Model 6

Model 6 includes CAR Normal spatially structured random effects for every disease with a CAR shared-parameter.

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{i,1}) &= \log(E_{i,1}) + \alpha_1 + \psi_i \times \gamma + \phi_{i,1} \\
\log(\mu_{i,2}) &= \log(E_{i,2}) + \alpha_2 + \psi_i / \gamma + \phi_{i,2} \\
\boldsymbol{\psi} &\sim \text{Normal}(\mathbf{0}, \sigma_{\psi}^2 (\boldsymbol{D} - \boldsymbol{W})^{-1}) \\
\boldsymbol{\phi}_j &\sim \text{Normal}(\mathbf{0}, \sigma_{\phi_j}^2 (\boldsymbol{D} - \boldsymbol{W})^{-1}) \\
\alpha_j &\sim \text{Normal}(0, 1000) \\
\sigma_{\psi} &\sim \text{Uniform}(0, 10) \\
\sigma_{\phi_j} &\sim \text{Uniform}(0, 10) \\
\gamma &\sim \text{Normal}(0, 100) \\
\boldsymbol{W}_{n \times n} &= \begin{cases} w_{ij} = 1 & \text{if } i \sim j \\ w_{ij} = 0 & \text{if } i = j \\ w_{ij} = 0 & \text{otherwise} \end{cases} \\
w_{i+} &= \sum_j w_{ij} \\
\boldsymbol{D}_{n \times n} &= \text{diagonal}(w_{1+}, \ldots, w_{n+})
\end{aligned}
$$

$\boldsymbol{\psi}$ are the spatially structured shared-parameter components, γ is a scaling parameter, $\boldsymbol{\phi}_j$ are spatially structured random effects, $\boldsymbol{W}$ is the proximity matrix, and $\sigma_{\phi_j}^2$, σ_{ψ}^2 are variance parameters for $\boldsymbol{\phi}_j$ and $\boldsymbol{\psi}$.

Models 7 and 8

Models 7 is the Generalized Multivariate CAR model [3], where the CAR Normal spatially structured random effects of ZVD by small area are conditioned by the CAR Normal spatially structured random effects of dengue. Model 8 presents the Generalized Multivariate CAR model, where the CAR Normal spatially structured random effects of dengue per

[3] Jin X, Bradley P. Carlin BP, and Banerjee S. Generalized Hierarchical Multivariate CAR Models for Areal Data. Biometrics. 2005; 61(4): 950-961.

area are conditioned by the CAR Normal spatially structured random effects of ZVD.

$$
\begin{aligned}
O_{ij} &\sim \text{Poisson}(\mu_{ij}) \\
\log(\mu_{ij}) &= \log(E_{ij}) + \phi_{ij} \\
\boldsymbol{\phi}_1|\boldsymbol{\phi}_2 &\sim \text{Normal}(\boldsymbol{\delta}_1, \boldsymbol{\Xi}_1) \\
\boldsymbol{\delta}_1 &= \beta_1\mathbf{1} + (\eta_0\mathbf{I} + \eta_1\boldsymbol{W})(\boldsymbol{\phi}_2 - \beta_2\mathbf{1}) \\
\boldsymbol{\Xi}_1 &= \sigma_1^2(\boldsymbol{D} - \kappa_1\boldsymbol{W})^{-1} \\
\boldsymbol{\phi}_2 &\sim \text{Normal}(\boldsymbol{\delta}_2, \boldsymbol{\Xi}_2) \\
\boldsymbol{\delta}_2 &= \beta_2\mathbf{1} \\
\boldsymbol{\Xi}_2 &= \sigma_1^2(\boldsymbol{D} - \kappa_2\boldsymbol{W})^{-1} \\
\beta_j &\sim \text{Normal}(0, 100) \\
\eta_0, \eta_1 &\sim \text{Normal}(0, 10) \\
\sigma_j &\sim \text{Uniform}(0, 10) \\
\kappa_j &\sim \text{Uniform}(0, 0.99) \\
\boldsymbol{W}_{n\times n} &= \begin{cases} w_{ij} = 1 & \text{if } i \sim j \\ w_{ij} = 0 & \text{if } i = j \\ w_{ij} = 0 & \text{otherwise} \end{cases} \\
w_{i+} &= \sum_j w_{ij} \\
\boldsymbol{D}_{n\times n} &= \text{diagonal}(w_{1+}, \ldots, w_{n+})
\end{aligned}
$$

where $\boldsymbol{\phi}_j$ are spatially structured random effects, $\boldsymbol{W}$ is the proximity matrix, $\boldsymbol{\phi}_1|\boldsymbol{\phi}_2$ is the conditional distribution of $\boldsymbol{\phi}_1$, and $\boldsymbol{\phi}_2$ is a marginal distribution.

7.6.1 Model goodness of fit

For model diagnosis we use posterior predictive checks, residual histograms, and scatter plots and Spearman correlation coefficient of the observed versus the fitted counts. To assess whether the model recovers the over dispersion observed in the data we compare the ratio $C_Z = Var(Z)/\bar{Z}$ based on sampled new data with the corresponding ratio C_y for the observed counts[4]. This check would be done at each iteration and a satisfactory model will have C_Z exceeding C_y about 50% of the time.

$$P_c = E[Pr(C_z > C_y|y)]$$

Values of $\hat{P}_c$ near 0 or 1 (above 0.9 or below 0.1) indicate discrepancy between the observations and the model, while values close to 0.5 mean that the observed data and the fitted data sampled from the model are closely comparable in terms of the overdispersion function.

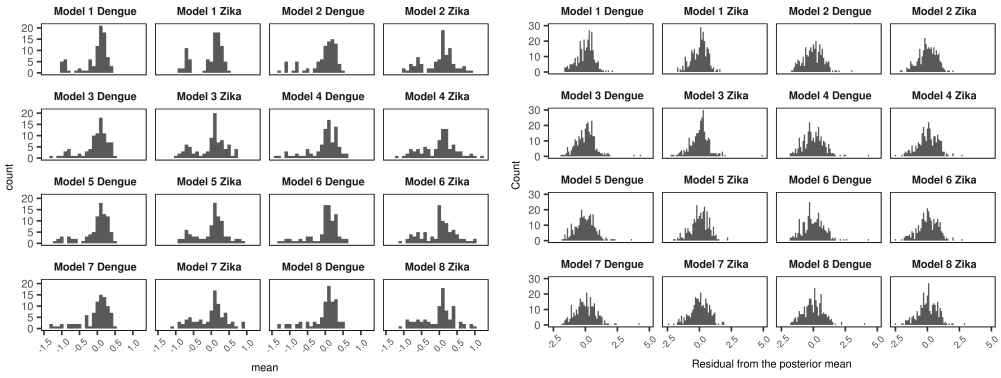
[4]Congdon, P. Bayesian models for categorical data. John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex PO19 8SQ, England. 2005. 447 pages.

Then, Table 7.6 shows the predictive checks on overdispersion from the joint models 1 to 8 in Santander and Bucaramanga. Following the decision criteria, all the joint models for the deparment of Santander recovers the overdispersed data, because the predictive checks show an average probability of $P_c = 0.50$, while the predictive checks for the city of Bucaramanga show some models with acceptable overdispersion recovery (models 1 to 7 for dengue, and models 1, 3 and 8 for ZVD), and with acceptable but close to the boundary probabilities for model 8 for dengue and models 2, 4, 5,6, and 7 for ZVD. The conclusion is that in the city of Bucaramanga all models clearly recovered the overdispersion, however ZVD data were more difficult to fit than dengue data, and in the department of Santander, all models recovered overdispersion very well.
Figure 7.6 displays the residual histograms of the posterior mean of the fitted values. At departmental level, model residuals shrinks to zero, and the residual boundaries are -1.5 to 1. In the city of Bucaramanga, model residuals are more dispersed than residuals in Santander, and the residual boundaries are from -2.5 to 5.0. From the residual examination, similar to the predictive checks in overdispersion, we observe that the joint models fitted the data worse for the city level than for the departmental level.

Table 7.6: Posterior predictive check for overdispersion, joint models of the city of Bucaramanga and the department of Santander.

Santander			Bucaramanga		
Model	Dengue	Zika	Model	Dengue	Zika
1	0.50	0.50	1	0.71	0.78
2	0.51	0.50	2	0.51	0.88
3	0.51	0.49	3	0.62	0.68
4	0.50	0.49	4	0.43	0.81
5	0.51	0.50	5	0.70	0.83
6	0.51	0.50	6	0.41	0.82
7	0.51	0.50	7	0.39	0.86
8	0.49	0.50	8	0.81	0.48

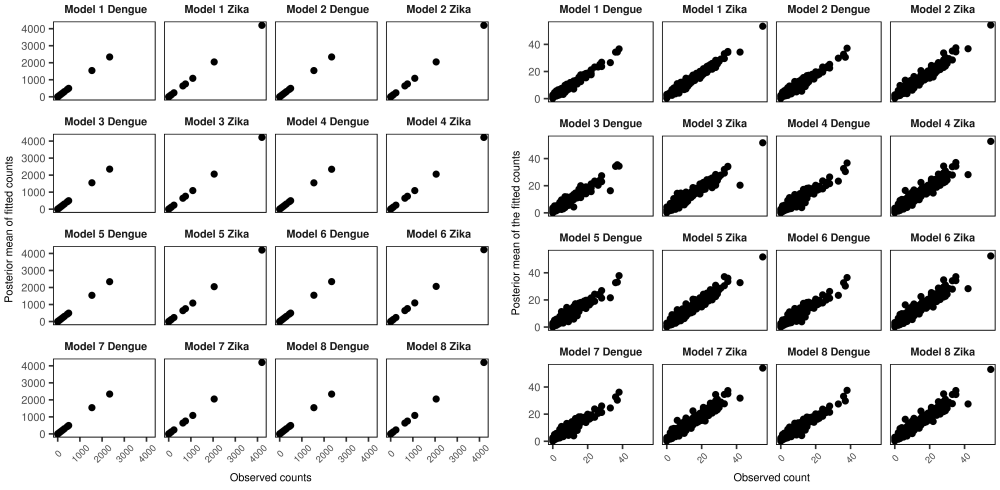
The association between the posterior mean of the fitted values and the observed counts of dengue or ZVD is displayed using scatter plots of fitted versus observed and the assessment of the Spearman correlation coefficient. For the department of Santander, the association of fitted and observed values in Figure 7.7 reveals that all models generate fitted values very close to the observed counts, which is ratified by the near to one Spearman correlation coefficients from Table 7.7. For the city of Bucaramanga, we observe that there are some variability in the association fitted-observed, so the models are missing something in following the observed counts, which is confirmed by the Spearman correlation coefficients near to 0.950 for most of the models.
As an overall conclusion, joint models work very well by predicting the observed counts of dengue and ZVD, recovering the data overdispersion, and reducing the dispersion revealed in the residuals at the departmental level, although the performance could be improved at the city level.
Reasons for the lower performance of the joint models at city level could be explained by



(a) Department of Santander (b) City of Bucaramanga.

Figure 7.6: Histograms of the residuals from the posterior mean of the fitted counts obtained from the joint models 1 to 8, department of Santander , and city of Bucaramanga.

the need to include other parameters in the linear predictor, for instance the inclusion of correlated as well uncorrelated heterogeneity parameters at the same time, or perhaps to difficulties of the model to fit high number of areas (293 at city level compared with 87 areas at departmental level).
Although the lower performance of the joint models at city level is not an ideal predictive feature, we have observed that the linear association fitted-observed counts is greater than 0.92, which is not so bad, so we can proceed to make model selection based on the DIC.



(a) Department of Santander. (b) City of Bucaramanga.

Figure 7.7: Scatter plots of the posterior mean of fitted counts versus observed counts obtained from the joint models 1 to 8.

Table 7.7: Spearman correlation coefficient between the posterior mean of the fitted counts and the observed counts, joint models 1 to 8 of the department of Santander and the city of Bucaramanga.

Santander			Bucaramanga		
Model	Dengue	Zika	Model	Dengue	Zika
1	0.995	0.994	1	0.962	0.972
2	0.995	0.987	2	0.938	0.958
3	0.995	0.992	3	0.944	0.957
4	0.992	0.976	4	0.925	0.945
5	0.995	0.984	5	0.932	0.949
6	0.994	0.980	6	0.922	0.945
7	0.994	0.988	7	0.928	0.952
8	0.995	0.984	8	0.929	0.947

7.6.2 Calculating expected values of dengue and ZVD

The joint models of relative risk require the observed counts and the expected counts of the disease in the *i*th area. Expected counts are obtained by external or internal standardization. External standardization requires incidence rates (cases per 100,000 people) by age groups and sex obtained from a standard or reference population, while internal standardization uses incidence rates obtained from the same data. In the manuscript, we used internal standardization to compute the expected values. The incidence rate is computed by

$$\text{IR}_{pq} = \frac{\text{Cases}_{pq}}{\text{Population}_{pq}} \times 100,000$$

where IR_{kl} is the incidence rate in age groups $p = 1, \ldots, P$, $(1 = [0,5), 2 = [5,10), \ldots, 14 = [65,100))$ and sex $q = 1,2$, (1=female, 2= male); Cases_{pq} are the total number of ZVD or dengue cases; and Population_{pq} is the total population in age group p and sex q for the departmental or city level over the complete study period, as obtained from the 2016 projected population of the Colombian census 2005 Using the IR_{pq}, the expected values of dengue or ZVD per municipality at departmental level, or census section at city level were calculated using

$$E_i = \sum_{p=1}^{P} \sum_{q=1}^{2} (\text{IR}_{pq} \times \text{Population}_{ipq})$$

where E_i is the expected value in small area i, and Population_{ipq} is the census population in small area i, age group p and sex q. Finally, four sets of expected values were obtained: Santander's ZVD and dengue expected values, and Bucaramanga's ZVD and dengue expected values.
The ratio between the observed and the expected values of dengue and ZVD per area i is the standardized incidence ratio (SIR). The SIR is a raw estimate of the disease risk.

$$\text{SIR}_i = \frac{\text{Observed}_i}{\text{Expected}_i}$$

Chapter 8

Real time parameter estimation of Zika outbreaks using model averaging

Abstract

Early prediction of the final size of any epidemic and in particular for Zika disease outbreaks can be useful for health authorities in order to plan the response to the outbreak. The Richards model is often been used to estimate epidemiological parameters for arboviral diseases based on the reported cumulative cases in single and multiwave outbreaks. However, other nonlinear models can also fit the data as well. Typically, one follows the so called post selection estimation procedure, i.e., selects the best fitting model out of the set of candidate models and ignores the model uncertainty in both estimation and inference since these procedures are based on a single model. In this paper we focus on the estimation of the final size and the turning point of the epidemic and conduct a real-time prediction for the final size of the outbreak using several nonlinear models in which these parameters are estimated via model averaging. The proposed method is applied to Zika outbreak data in four cities from Colombia, during the outbreak ocurred in 2015-2016.

Keywords: 3-parameter logistic, 4-parameter Gompertz, 5-parameter logistic, Richards, Weibull.

8.1 Introduction

Zika infection is an arboviral disease characterized by sub-clinical or mild dengue-like illness, with severe manifestations such as Guillain-Barre syndrome in adults and microcephaly in babies born to infected mothers [1]. During 2015 and 2016 the disease affected several south-American countries, especially Brazil and Colombia. In Colombia, a total number of 104755 cases were recorded, from which 8826, 92113 and 3816 cases were confirmed by laboratory, diagnosed by clinical symptoms and suspected without confirmation, respectively. Zika cases were recorded between the 32th epidemiological week, 2015 until the 42nd epidemiological week, 2016 (09/08/2015-22/10/2016) [2]. For the analysis presented in this paper we use the data from four cities, Bucaramanga (3651 cases), Cali (12220), Cúcuta (4287) and Neiva (1940).

Mathematical and statistical models are increasingly being used to facilitate the estimation of the primary epidemiological parameters in infectious disease outbreak. During a single peak outbreak, the turning point (or primary inflection point), i.e. the point in time at which the rate of accumulation changes from increasing to decreasing, and the final size of the epidemic are among the epidemiological parameters of interest to be estimated [3] [4] [5]. Once an outbreak has begun, knowledge about the potential severity in real-time (i.e., before the end of the outbreak) can help public health authorities to respond effectively [5].

Various epidemiological studies used mathematical and statistical models to describe the evolution and spread of severe acute respiratory syndrome (SARS) and dengue and to evaluate the impact of control interventions. In particular, Hsieh *et al.* [3] [5] and Zhou *et al.* [4] proposed to use a nonlinear model, the Richards model [6], in order to model the cumulative number of reported cases, to estimate the turning point and the basic reproduction number, R_0. Hsieh *et al.* applied nonlinear models to model epidemics of SARS [7], dengue disease [8] [9] [10], Influenza A (H1N1) [11] and ebola [12]. In addition to the estimation of the epidemiological parameters, Hsieh *et al.* [3] used the model for real-time prediction of these parameters for dengue and SARS outbreaks.

Hsieh *et al.* estimation and prediction is based on the Richards model and considers only the cumulative infective population size with saturation in growth as the outbreak progresses. The basic premise of the Richards model is that the incidence curve consists of a single peak of high incidence, resulting in an S-shaped epidemic curve and a single turning point of the outbreak [7] [8]. Hsieh *et al.* [7] [8] also showed that the Richards model can be used to model multiwave outbreaks as well. In this paper we focused on a single wave outbreak.

A variety of nonlinear models have been developed to model growth data. Among them, we consider the 3-parameter logistic (3P logistic) [7] [13] [14], 5-parameter logistic (5P logistic) [15], Sigmoid Emax [16], Gompertz [14] [17] and Weibull [18] models. All these models can be used to fit epidemic data as well. Fitting several models to the same data raises the issue, central in statistical modeling, of model selection. Indeed, a model selection procedure is needed in order to choose the model with the best fit to the data. Often, one is confronted with the problem that several models are performing

almost equally well over the range of observed data. Typically, one selects the best fitting model out of the set of fitted models and ignores the uncertainty due to model selection in estimation and inference. For these reasons, several authors (i.e., Burnham & Anderson [19], Claeskens & Hjort [20], Posada [21], Moon [22] and [23]), advocate the use of model averaging techniques to perform multi-model estimation and inference. Model averaging is a method that takes into account all fitted models for the estimation of the parameters of primary interest. It is based upon a weighted average of the parameter of primary interest obtained from different models, giving largest weights to those models that best fit the data [24].
In the current paper we analyze Zika outbreak data and estimate a model average of the final size and the turning point of the epidemic, and perform a real-time prediction using several nonlinear models. A real-time prediction is a procedure in which the final size of the outbreak is estimated as early as possible. An elaborate description of the procedure is given in the Supplementary Material for the paper (Figure 8.4). The proposed method is applied to four Zika outbreaks that occurred in four cities in Colombia during the 2015/2016 outbreak.

8.2 Data

The data used in this paper were collected from four cities from Colombia, where Zika disease cases were reported to the Instituto Nacional de Salud (Colombian National Institute of Health) in a weekly basis. The study locations represent cities with the highest number of Zika cases within all the cities in Colombia. Two of the cities are located to the northeast of Colombia (Bucaramanga and Cúcuta) and two are located to the southwest (Cali, Neiva).
The weekly counts of clinical Zika cases by date of onset of symptom per city were converted into cumulative case curves starting on the 50 epidemiologic week (EW) of 2015 for Cali and Bucaramanga, and the 47 EW for Cucuta and Neiva. The starting and ending dates of the outbreaks, as well as the observed number of cases at the end and the observed attack rates are shown in Table 8.1.
Figure 8.1 shows incidence and cumulative number of cases for the four cities under study. The outbreaks in Cúcuta and Neiva started and ended earlier (47 EW 2015 until 13 EW 2016) than the outbreaks in Cali and Bucaramanga (50 EW 2015 until 36 EW 2016).

8.3 Methods

8.3.1 Modeling Zika outbreak using nonlinear models

The Richards model [6], [25], [26] has often been used to model reported cumulative cases in disease outbreaks. In particular, Hsieh *et al.* [5] [7] demonstrated that Richards model can be used for real-time prediction of outbreak severity by estimating the carrying

Table 8.1: Epidemiological information on the 2015/2016 Zika outbreak in the four cities from Colombia.

Area	Bucaramanga	Cali	Cúcuta	Neiva
Population	528,269	2,394,925	656,380	344,026
Elevation (meters)	959	1000	320	442
Start (2015)	13 December	13 December	22 November	22 November
End (2016)	10 September	10 September	02 April	02 April
Outbreak length (EW)	39	39	19	19
Observed final size	3,651	12,220	4,287	1,940
Cases $\times$ 100,000	69.11	51.02	65.31	56.39

capacity (i.e. the final size of the epidemic) as well as real-time detection of the turning point (i.e. the time point with the peak number of cases) of the epidemic.
In Hsieh *et al.* [5] [7] [8] [9] [10], the estimates were obtained under the assumption that the cumulative number of reported cases at time t, Y_t, are normally distributed with mean $\mu(t,\theta)$ and variance σ^2, $Y_t \sim N(\mu(t,\theta),\sigma^2)$. The mean structure for $\mu(t,\theta)$ is given in the first line in Table 8.2. As pointed out by Hsieh *et al.* [5] [7] [8] [9] the parameter vector to be estimated is $\theta = (\alpha,\gamma,k,\eta)$, where α is the final size of the epidemic, γ is the per capita intrinsic growth rate of the infected population, k is the exponent of the deviation from the standard logistic curve and η is the turning point.
The cumulative number of reported cases in a Zika outbreak is an example of growth data. For many types of growth data, the growth rate does not steadily decline, but rather increases to a maximum before steadily declining to zero. In such models, η is the position of the point of inflection (turning point of the epidemic for the application presented in this paper), the time when the growth rate is greatest.
Table 8.2 presents other five possible nonlinear models. Note that all these models are scaled in calendar time (for which $t = 1$ is the first epidemiological week in which the outbreak ocurred). The three-parameter logistic model [13] [14] (3P logistic) is a special case of the Richards model, obtained when the exponent $k = 1$. For the 3P logistic model, the growth curve is symmetric around turning point and has equal periods of slow and fast growth. The Gompertz model [17] [14] is another special case of the Richards function when $\gamma \to 0$ and is frequently used in situations where growth is not symmetrical about the turning point. There are many variants of the Weibull model, the one we use in this paper is a modification of the Gompertz model when its independent variable, time, is rescaled by logarithmic transformation [18]. Note that for all the models in Table 8.2 the turning point and the final size of the epidemic are parameters in the model.
The sigmoid Emax model and the five-parameter logistic (5P logistic) are commonly used in dose-response modeling [23]. The sigmoid Emax model [16] is obtained by mathematical transformation of 3P logistic model and rescaling the independent variable by a logarithmic transformation. Similarly, the 5P logistic model is obtained by rescaling the independent variable by logarithmic transformation and by doing a reparametrization,

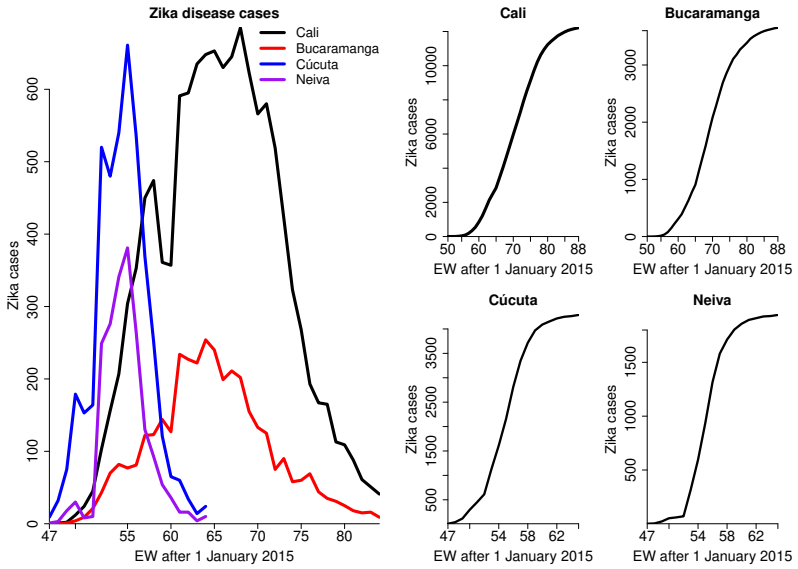


Figure 8.1: Weekly number of cases (left) and cumulative cases (right) of Zika disease for the 2015/2016 outbreak in four cities from Colombia.The time scale is given in epidemiological weeks (EW).

so that the model evaluated at the inflection point (η) reaches 50% of maximum response [15].
The first derivative of $\mu(t,\theta)$, $\mu^{'}(t,\theta) = \frac{\partial \mu(t,\theta)}{\partial t}$, is the number of cases at time t, i.e., the incidence at time t. Except for the 3P logistic model, $\mu^{'}(t,\theta)$ is not symmetric around the turning point.

8.3.2 Model uncertainty, model selection and model averaging

In the previous section, we presented six nonlinear models that can be used for the estimation of epidemiological parameters and for real time prediction. In this section, we describe the model averaging (MA) technique (Burnham & Anderson [19] [27], Claeskens & Hjort [20]), which is used to account for model uncertainty by combining together the estimates from all the fitted models. Within the model averaging framework, one fits a set of R candidate models, $g_1, g_2, \ldots, g_R$, to the data in order to obtain the parameter estimates from all models, $\widehat{\theta}_1, \widehat{\theta}_2, \ldots, \widehat{\theta}_R$. A post selection procedure [20] [23] implies that we first need to select the model with the best goodness-to-fit to the data, say g_ℓ, and to estimate θ by $\widehat{\theta}_\ell$. The model selection can be based on an information criteria. However, this procedure does not take into account model uncertainty since the estimation is based on a single model. The model averaging techniques allow us to estimate the component in θ using information obtained from all fitted models and in that way to account for model uncertainty. Let us assume that the Akaike's Information Criterion

Table 8.2: Nonlinear models considered to fit the cumulative cases of Zika outbreak.

Models	$\mu(t,\theta)$	$\mu^{'}(t,\theta)$
Richards	$\frac{\alpha}{\left[1+k.e^{-\gamma(t-\eta)}\right]^{\frac{1}{k}}}$	$\gamma\mu(t)\left[1-\left(\frac{\mu(t)}{\alpha}\right)^{k}\right]$
3P logistic	$\frac{\alpha}{1+e^{-\gamma(t-\eta)}}$	$\gamma\mu(t)\left[1-\frac{\mu(t)}{\alpha}\right]$
5P logistic	$\alpha+\frac{\alpha_0-\alpha}{[1+(2^{\frac{1}{k}}-1)(\frac{t}{\eta})^{\gamma}]^{k}}$	$-\frac{k\gamma}{t}[\mu(t)-\alpha][1-(\frac{\mu(t)-\alpha}{\alpha_0-\alpha})^{\frac{1}{k}}]$
Sigmoid Emax	$\alpha_0+\frac{t^n(\alpha-\alpha_0)}{t^n+\eta^n}$	$\frac{n}{t}[\mu(t)-\alpha_0][1-\frac{\mu(t)-\alpha_0}{\alpha-\alpha_0}]$
4P Gompertz	$\alpha_0+(\alpha-\alpha_0)e^{-e^{-\gamma(t-\eta)}}$	$-\gamma[\mu(t)-\alpha_0]ln[\frac{\mu(t)-\alpha_0}{\alpha-\alpha_0}]$
Weibull	$\alpha+(\alpha_0-\alpha)e^{-(\frac{t}{\eta})^{\gamma}}$	$\frac{\gamma}{t}[\mu(t)-\alpha]ln[\frac{\mu(t)-\alpha}{\alpha_0-\alpha}]$

(*AIC* [28]) is used for model selection. For a given set of R candidate models, g_1, g_2,..., g_R, Burnham & Anderson [27] proposed to rescale the *AIC* to

$$\Delta AIC_i = AIC_i - AIC_{min}, i = 1,...,R$$

Here, AIC_{min} is the smallest *AIC* value across the set of R models. The *AIC* differences, ΔAIC_i, are interpreted as the information loss when model g_i, rather than the best model g_{min}, is used to approximate the true and unknown model. Burnham & Anderson [19] defined Akaike's weights as

$$w_i(AIC) = \frac{exp\left(-\frac{1}{2}\Delta AIC_i\right)}{\sum_{i=1}^{R} exp\left(-\frac{1}{2}\Delta AIC_i\right)}.$$

Akaike's weight $w_i(AIC)$ can be interpreted as the weight of evidence that model g_i is the best model given a set of R models and given that one of the models in the set must be the best model. The nonlinear model with the highest Akaike's weight (i.e, the minimum AIC) is considered as the model with the best goodness-to-fit to the data.
Following Burnham & Anderson [19], we can calculate the model averaged estimator for turning point ($\widehat{\eta}_{MA}$) and the final size of outbreak ($\widehat{\alpha}_{MA}$) as follow:

$$\widehat{\eta}_{MA} = \sum_{i=1}^{R} w_i(AIC)\,\widehat{\eta}_i,$$

$$\widehat{\alpha}_{MA} = \sum_{i=1}^{R} w_i(AIC)\,\widehat{\alpha}_i.$$

Here, $\widehat{\eta}_i$ and $\widehat{\alpha}_i$ are the parameter estimates for the turning point and final size of outbreak of ith model, respectively. The estimators for variance for $\widehat{\eta}_{MA}$ and $\widehat{\alpha}_{MA}$ are given, respectively, by:

$$\widehat{var}(\widehat{\eta}_{MA}) = \left[\sum_{i=1}^{R} w_i(AIC)\sqrt{\widehat{var}(\widehat{\eta}_i|M_i) + (\widehat{\eta}_i - \widehat{\eta}_{MA})^2}\right]^2,$$

$$\widehat{var}\left(\widehat{\alpha}_{MA}\right) = \left[\sum_{i=1}^{R} w_i\left(AIC\right)\sqrt{\widehat{var}\left(\widehat{\alpha}_i|M_i\right) + \left(\widehat{\alpha}_i - \widehat{\alpha}_{MA}\right)^2}\right]^2 .$$

Note that one can replace the AIC by other information criteria such as BIC, KIC and calculate the model's weight based on these criteria.
In addition, since the models' weights are based on the AIC (or any other information criterion) the model averaging approach described above ensure that the parameter estimates for the turning point and final size of the epidemic (and their standard errors) will be dominated by the model(s) with the best goodness-of-fit.

8.4 Results

8.4.1 Estimation of the final size and turning point using model averaging methods

All non-linear models discussed above were fitted to the single-phase Zika outbreak in the cities of Bucaramanga, Cali, Cúcuta and Neiva. The models were fitted to the weekly cumulative number of reported cases and the turning point and the final size were estimated for each city. The models were fitted using R software 3.3.1 [29], using the `gnls` function from package `nlme` [30].
Table 8.3 shows the parameter estimates, Akaike's information criteria and the Akaike's weights for all non-linear models used to calculate the model averaged estimates for the turning point and the final size of the outbreak. Figure 8.2 displays the cumulative predicted values and the incidence predicted values obtained for the fitted models to the complete data in each city, together with the observed values.
For Bucaramanga, the 5PL model ($AIC = 348.6$), the Richards model ($AIC = 361.9$) and the 3PL model ($AIC = 397.0$) had lower AIC than the 4P Gompertz model ($AIC = 424.1$), the sigmoid Emax model ($AIC = 405.5$) and the Weibull model ($AIC = 468.3$). The model averaging final size estimate of the Zika outbreak is equal to 3700 cases (95% CI: 3632, 3768) with a turning point close to the 19.1 weeks (95% CI: 18.8, 19.3) after the begining of the outbreak. Note that the observed final size is equal to 3651 (Table 8.1).The 5P logistic model has an Akaike weight equal to 0.999, while the weights obtained for the other models are relatively small which implies that the model average parameter estimated will be dominated by the 5P logistic model for this city.
Similar pattern was observed in Cali. The 5PL model ($AIC = 451.6$), the Richards model ($AIC = 494.9$) and the 3PL model ($AIC = 511.7$) had lower AIC than the 4P Gompertz model ($AIC = 527.0$), the sigmoid Emax ($AIC = 521.1$) and the Weibull model ($AIC = 567.2$). The final size estimate for the Zika epidemic from the model averaging is equal to 12458 cases (95% CI: 12023, 12893), with an estimated turning point of 20.3 weeks (95% CI: 19.9, 20.8) after the Zika epidemic starts. As for Bucaramanga, in Cali, the Akaike's weight from the 5P logistic model is the highest and closed to 1, which implies that model averaging parameter estimates are mostly based on the 5P logistic model. The 5P logistic model is the model with the smallest AIC ($AIC = 196.8$) in Cúcuta

as well. The 3P logistic (AIC =196.8) and Richards model (AIC =197.4) have similar AIC. Once again, the Weibull model has the highest information criteria (AIC =247.9). The model averaging estimate for final size is 4285 cases (95% CI: 4199, 4372), with an estimated turning point of 8.93 weeks (95% CI: 8.72, 9.15) after the outbreak begins. For this city Akaike's weights for the 3P logistic, the Richards and the 5P logistic models are equal to 0.370, 0.265 and 0.365, respectively. Hence, the model averaging estimates of final size and turning point are dominated by the parameter estimates obtained for these three models.
The model fit for the city of Neiva is sligthly different from the other cities reported above. Here, the models with the lowest AIC were the sigmoid Emax model ($AIC = 173.8$), the 5PL model ($AIC = 175.7$) and the Richards model ($AIC = 176.9$), while the model with the highest AIC corresponded again to the Weibull model ($AIC = 198.3$). The model averaging final size estimate for Neiva is equal to 1943 cases (95% CI: 1913, 1972), and an estimated turning point of 8.9 weeks (95% CI: 8.7, 9.3) after the outbreak starts. In Neiva, the highest Akaike's weight is for the sigmoid Emax model (0.613), followed by the 5P logistic (0.236) and Richards model (0.133).

8.4.2 Real-time prediction

The model average framework is particularly useful for real-time prediction since we use only part of the data for long term prediction. In this case, we do not want to base the estimation on a single model. Point estimates and 95% CI obtained for all nonlinear models and from the model averaging are presented in the supplementary appendix of the paper, for the turning point and final size of the outbreak in Bucaramanga, Cali, Cúcuta and Neiva (Supplementary Table 8.5, Table 8.6, Table 8.7 and Table 8.8, respectively).
For Bucaramanga, the model averaging point estimate for the final size of the epidemic is 3452 cases (95% CI: 2982, 3922), for the model fitted using the data of the first 26 weeks after the epidemic starts (Figure 8.3 d and Supplementary Table 8.5), while the observed final size of the outbreak is 3651 cases (Table 8.1). This implies that around 4 months before the end of the epidemic this valuable information could be available for the health authority. Furthermore, for the same estimation period, the model averaging estimate for the turning point becomes stable with a point estimate of 18.6 weeks (17.4,19.8) (Figure 8.3 b and Supplementary Table 8.5). These results indicate that the health authority could estimate the turning point of the outbreak about 6 weeks after its occurrence in Bucaramanga.
For Cali, the observed final size of the Zika epidemic is 12,220 cases (Table 8.1). All the models overestimate the final size of the epidemic. The model average estimate stabilizes around the estimate of 12,560 cases (95% CI: 11,848; 13,272) for the models fitted to the period 1 to 33 weeks after the epidemic starts (i.e, 6 weeks before the end of the epidemic Figure 8.3 h and Supplementary Table 8.6). The model averaging estimate for the turning point obtained from the model fitted to the estimation period 1 to 33 weeks is 22.5 weeks (95% CI: 21.5, 19.4) (Figure 8.3 f and Supplementary Table 8.6).

Table 8.3: Parameter estimates for the turning point and final size of the epidemic obtained for the six nonlinear model and their model average estimates per city.

City	Model	Turning Point	Final Size estimate	k	AIC	Weight
Bucaramanga	3P logistic	19.0 (18.9,19.1)	3623 (3596,3651)		397.0	2.997e-11
	Richards	18.3 (18.1,18.5)	3680 (3654,3705)	0.59 (0.48,0.69)	361.9	0.001
	5P logistic	19.1 (19.0,19.2)	3700 (3672,3729)		348.6	0.999
	Sigmoid Emax	19.3 (19.1,19.5)	3838 (3785,3892)		405.5	4.435e-13
	4P Gompertz	16.9 (16.8,17.2)	3789 (3732,3845)		424.1	4.055e-17
	Weibull	17.6 (17.1,18.0)	4206 (3994,4418)		468.3	9.976e-27
	Model Averaging	19.1 (18.8,19.3)	3700 (3632,3768)			
Cali	3P logistic	20.3 (20.1,20.5)	12350 (12204,12496)		511.7	8.751e-14
	Richards	19.5 (19.1,19.9)	12645 (12449,12841)	0.55 (0.38,0.72)	494.9	3.889e-10
	5P logistic	20.3 (20.2,20.4)	12458 (12333,12583)		451.6	1
	Sigmoid Emax	20.9 (20.6,21.3)	13510 (13166,13854)		521.1	7.846e-16
	4P Gompertz	18.1 (17.8,18.4)	13216 (12924,13508)		527.0	4.101e-17
	Weibull	19.5 (18.5,20.4)	15927 (14557,17298)		567.2	7.590e-26
	Model Averaging	20.3 (19.9,20.8)	12458 (12023,12893)			
Cúcuta	3P logistic	8.9 (8.8,8.9)	4293 (4255,4331)		196.8	0.370
	Richards	8.9 (8.8,9.2)	4281 (4237,4325)	1.13 (0.86,1.4)	197.4	0.265
	5P logistic	8.9 (8.9,9.0)	4280 (4231,4328)		196.8	0.365
	Sigmoid Emax	8.9 (8.8,9.2)	4433 (4313,4554)		226.9	1.019e-07
	4P Gompertz	8.0 (7.8,8.2)	4402 (4277,4528)		233.5	4.024e-09
	Weibull	8.2 (7.8,8.5)	4673 (4356,4991)		247.9	3.010e-12
	Model Averaging	8.9 (8.7,9.1)	4285 (4199,4372)			
Neiva	3P logistic	9.1 (8.9,9.2)	1912 (1888,1936)		184.7	0.003
	Richards	8.7 (8.5,8.9)	1929 (1905,1952)	0.49 (0.21,0.78)	176.9	0.133
	5P logistic	9.1 (8.9,9.1)	1943 (1916,1971)		175.7	0.236
	Sigmoid Emax	9.0 (8.9,9.1)	1945 (1924,1967)		173.8	0.613
	4P Gompertz	8.3 (8.2,8.4)	1946 (1921,1972)		181.3	0.014
	Weibull	8.3 (8.2,8.5)	2008 (1952,2064)		198.3	2.884e-06
	Model Averaging	8.9 (8.7,9.3)	1943 (1913,1972)			

For Cúcuta, The final size estimate from model averaging is stable from the model fitted for the estimation period 1 to 14 week, with a point estimate of 4308 cases (95% CI: 4039, 4578) (Figure 8.3 l and Supplementary Table 8.7). The observed final size for Cúcuta is 4287 cases (Table 8.1). The turning point estimate from model averaging is equal to 8.9 weeks (95% CI: 8.6, 9.2), for the models fitted to the period 1 to 14 weeks after the epidemic starts (i.e, 5 weeks before the end of the outbreak Figure 8.3 j and Supplementary Table 8.7).

For Neiva, the observed final size of the epidemic is 1940 cases (Table 8.1). The model average estimate for the final size of the epidemic is equal to 1899 cases (95% CI: 1742, 2055) obtained for the model fitted to the data of the first 13 weeks of the outbreak (i.e, 6 weeks before the end of the outbreak Figure 8.3 p and Supplementary Table 8.8). The turning point estimate from model averaging becomes stable for the models fitted to period 1 to 13 week, showing a point estimate of 8.9 weeks (95% CI: 8.5, 9.3) (Figure 8.3

n and Supplementary Table 8.8).
In general, the models seem to fit the data for the shorter outbreaks (Cúcuta and Neiva) better than for the longer outbreaks (Bucaramanga and Cali). In addition, the Weibull model fits the data poorly for all cities. An elaborate discussion about the performance of the Weibull model in Bucaramanga and Cali is given in the supplementary material of the manuscript. The 95% CIs for the final size of the epidemic for all cities contain the observed final size, which implies that around 2-4 months before the end of the outbreak an accurate estimate for the outbreak's final size could be provided to the health authorities. The turning point estimates for Cúcuta and Neiva underestimate the observed value by one week, but for Bucaramanga and Cali, the estimates are in line with the observed values, while the final size estimates estimate accurately the observed values for all cities (Figure 8.3). The model averaging estimates for the turning point and final size of the epidemic were stable around the first 2/3 of the total outbreak duration.

8.5 Discussion

Modeling a single wave outbreak requires to use a nonlinear growth model in order to estimate the epidemiological parameters of interest. In this study we have shown that several nonlinear models, the Richards, the 3PL, 5PL, the Weibull, sigmoid Emax and 4P Gompertz models, can be used to model the data. Alternatively, a model average technique that used a weighted parameter estimate, based on the model posterior probability, can be applied. In this paper we advocate the use of the model averaging technique since it does not ignore uncertainty related to model selection which is ignored when post selection inference and estimation is conducted.
Further, we have shown that the model averaging approach can be used in order to perform a real-time estimation for the turning point and prediction for the final size of the epidemic. We have shown that in the case of the Zika outbreak in 2015/2016 in Bucaramanga and Cali, an estimate with 95% CI that cover the observed final size could be given to the health authorities 4 months before the end of the epidemic, and for Cúcuta and Neiva accurate estimates for the final size could be provided 2 months before the end of the epidemic. Further, for a real time prediction, the model averaging offers an attractive modeling approach since the data available for modeling represents only a part of the outbreak data (from the start to the time in which the real-time prediction is performed). Hence, to our opinion, taking into account several possible models is appropriate in this setting.
Since this study was conducted retrospectively, with data from routine surveillance system, potential biases could not be prevented. Some limitations include data quality associated with real-time modeling (as data are often subject to ongoing cleaning, correction, and reclassification of onset dates as further data become available) and reporting delays. Whether reporting delays or dates of reporting and date of onset were known, it would be possible to perform more realistic analyses that include only cases known about at the end of the most recent time period. This would likely make the models appear less

attractive, but might provide a more realistic lower bound in terms of how quickly turning points can be identified.

Predicting the trend of an epidemic from limited data during early stages of the epidemic can be sometimes misleading. Nevertheless, early prediction of the magnitude of an epidemic outbreak is more important than retrospective studies [3]. The methodology proposed in the paper does not allow for a prediction of the turning point but only for the estimation. This is due to the fact that the nonlinear models presented in the paper fit poorly the data in the initial stage of the outbreak, i.e., when the turning point of the outbreak can be predicted. As was observed in other attempts at real-time prediction, the forecast appears to be very vulnerable to the timing of predictions, especially during the early phase of epidemic [31]. As long as the data include this inflection point and a time interval shortly after, the curve fitting and predicting future case number will be reasonably accurate [3].

The model averaging modeling approach provides an attractive framework for real-time prediction since it takes into account a set of models and the real-time prediction is dominated by the model(s) with the best goodness-of-fit to the data. In the case of Bucaramanga and Cali, one model dominated the estimation and prediction (the 5PL) while in the case of Neiva and Cucuta the model average estimates was dominated by several models (3PL, Richard and 5PL in Cucuta with a combined weight greater than 0.999 and Richard, 5PL, and sigmoid Emax with a combined weight of 0.982 in Neiva). This is the main advantage to use the model averaging framework for real-time prediction since, taking into account that only a part of the data is available and can be used for prediction, the estimation procedure is based on the best fitted models.

The present study successfully offers a modeling strategy to implement real-time prediction of an epidemic in the midst of its course. The methodology discussed in this paper was developed for a single wave outbreak. In future research we will extend the model averaging approach to a multi-wave outbreaks setting as well.

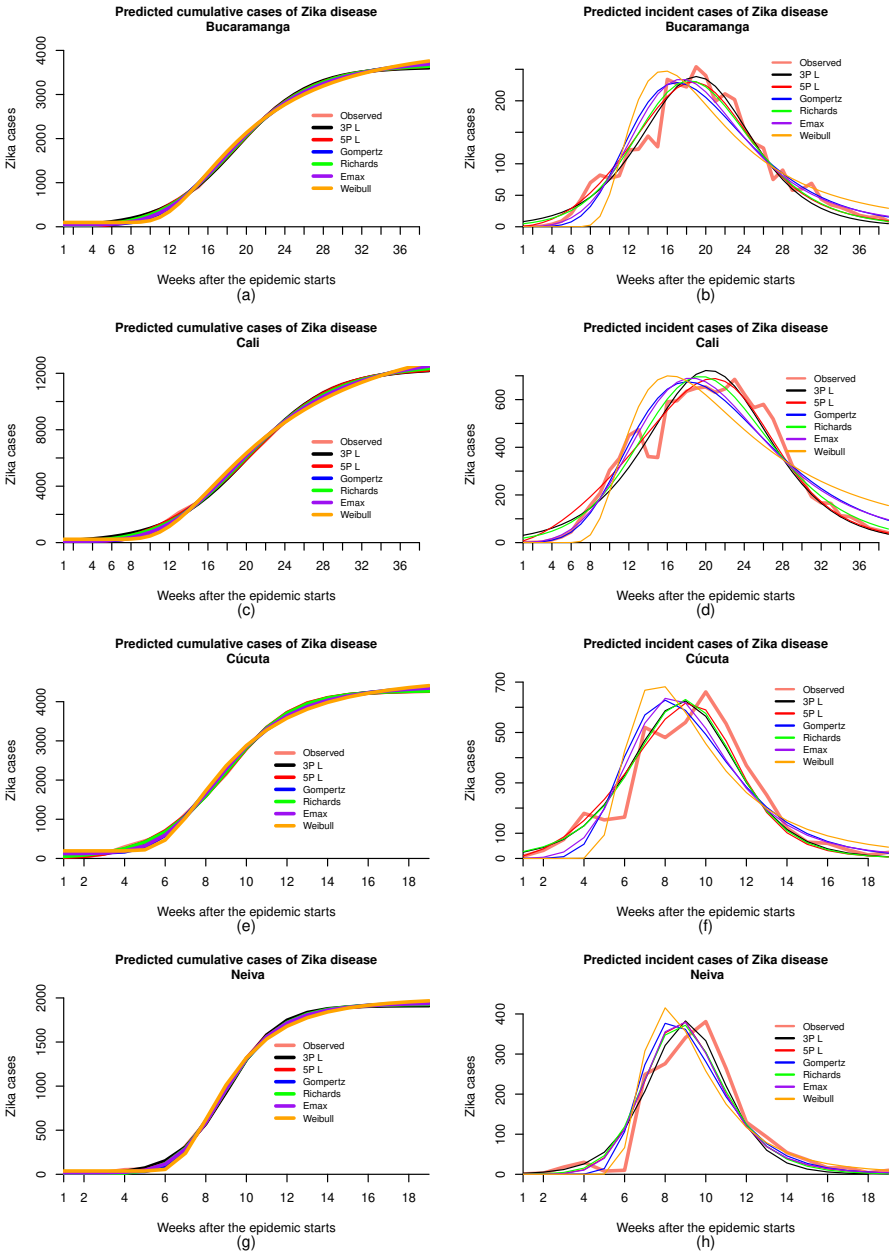


Figure 8.2: Predicted cumulative and incidence cases based on 6 nonlinear models for Zika outbreaks in four Colombian cities. Prediction is done when all data are used for the estimation of model parameters.

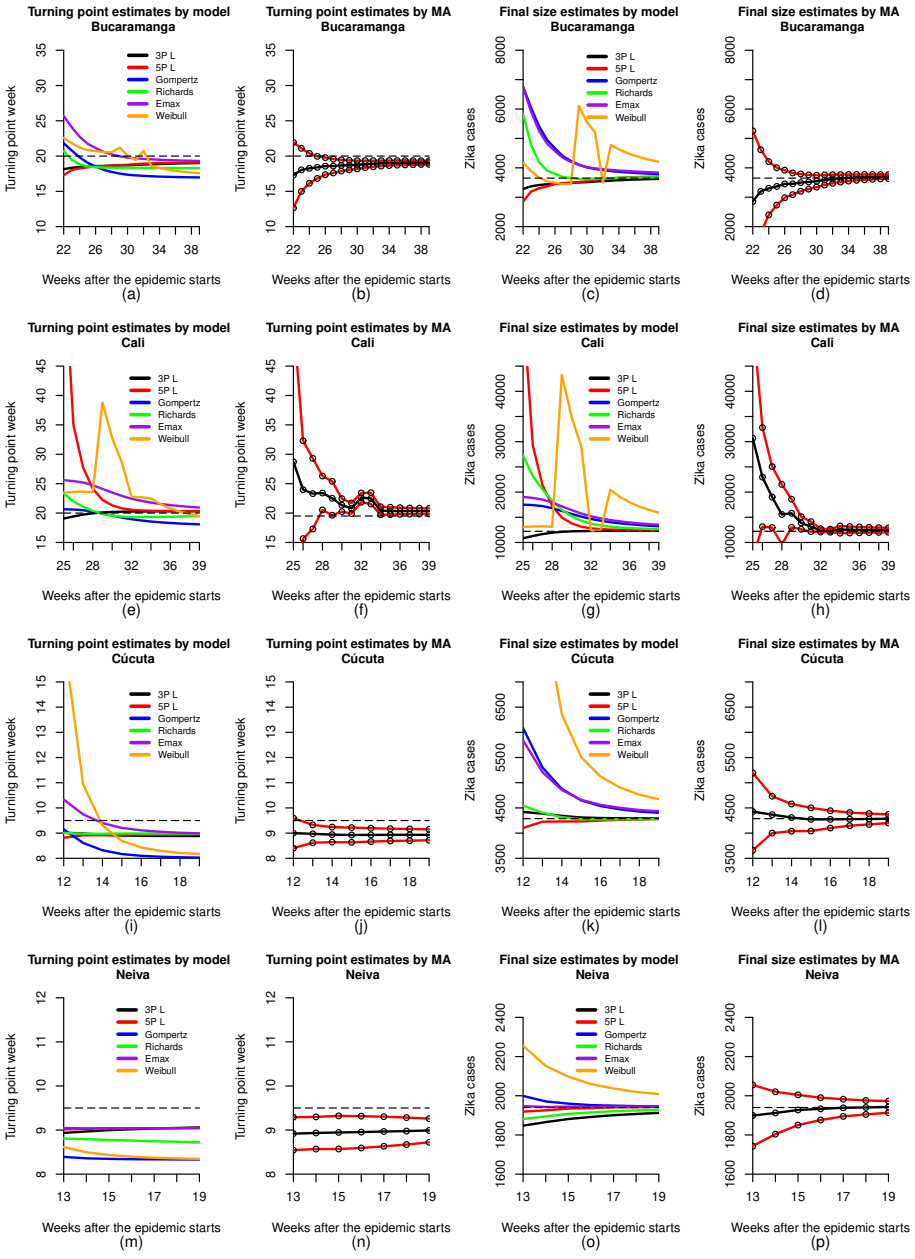


Figure 8.3: Parameter estimates for the turning point and final size of the outbreak, from the nonlinear models under study (point estimates), and from model averaging (MA) (point estimates and 95% CI) per city. Dashed lines represent the observed values. The time scale in all figures present the last week in the estimation period. For example in panel a, 22 implies that the estimation period is 1-22 weeks, etc.

8.6 References

[1] Plourde AR, Bloch EM. A Literature Review of Zika Virus. *Emerging Infectious Diseases* 2016; **22**: 1185–1192.

[2] Instituto Nacional de Salud. Reporte de notificación casos de Zika semana 32-2015 a semana 42-2016 [Zika case notification report, week 32 2015 - week 42 2016], Instituto Nacional de Salud [National Institute of Health], Colombia 2016. http://www.ins.gov.co/Noticias/ZIKA/Forms/AllItems.aspx

[3] Hsieh T, Lee J, Chang H. SARS epidemiology modeling. *Emerging Infectious Diseases* 2004; **10**: 1165–1167.

[4] Zhou G, Yan G. Severe Acute Respiratory Syndrome Epidemic in Asia. *Emerging Infectious Diseases* 2003; **9**: 1608–1610.

[5] Hsieh Y, Cheng Y. Real-time forecast of multi-wave epidemic outbreaks. *Emerging Infectious Diseases* 2006; **12**: 122–127.

[6] Richards F. A flexible growth function for empirical use. *Journal of Experimental Botany* 1959; **10**: 290–301.

[7] Hsieh Y. Richards model: A simple procedure for real-time prediction of outbreak severity, in: Z. Ma, J. Wu, Y. Zhoue (Eds.), *Modeling and Dynamics of Infectious Diseases* (Volumen 11), Contemporary Applied Mathematics (CAM), Higher Education Press, 2009; pp. 216–236. doi:10.1142/9789814261265_0009. http://mail.cmu.edu.tw/~hsieh/pdf/pub57.pdf

[8] Hsieh Y, Ma S. Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005. *American Journal of Tropical Medicine and Hygiene* 2009; **80**: 66–71.

[9] Hsieh Y, Chen C. Turning points, reproduction number, and impact of climatological events for multi-wave dengue outbreaks. *Tropical Medicine and International Health* 2009; **14**: 628–638.

[10] Hsieh Y, Arazoza H, Lounes R. Temporal trends and regional variability of 2001-2002 multiwave denv-3 epidemic in havana city: did hurricane michelle contribute to its severity? *Tropical Medicine and International Health* 2013; **18**: 830–838.

[11] Hsieh Y-H, Ma S, Velasco Hernandez JX, Lee VJ, Lim WY . Early Outbreak of 2009 Influenza A (H1N1) in Mexico Prior to Identification of pH1N1 Virus. *PLoS ONE* 2011; **6**:e23853 **10**:e0140810. doi:10.1371/journal.pone.0023853.

[12] Hsieh Y-H. Temporal Course of 2014 Ebola Virus Disease (EVD) Outbreak in West Africa Elucidated through Morbidity and Mortality Data: A Tale of Three Countries. *PLoS ONE* 2015; **10**:e0140810. doi:10.1371/journal.pone.0140810.

[13] Rozema E. Epidemic models for sars and measles. *The College Mathematics Journal* 2007; **38**: 246–259.

[14] Tsoularis A, Wallace J. Analysis of logistic growth models. *Mathematical Biosciences* 2002; **179**: 21–55.

[15] Liao J, Liu R. Re-parameterization of five-parameter logistic function. *Journal of Chemometrics* 2009; **23**: 248–253.

[16] MacDougall J. Analysis of dose-response studies-Emax model, in: N. Ting (Ed.), *Dose Finding in Drug Development*, Statistics for Biology and Health, Springer New York, 2006, pp. 127–145.

[17] Wellock I, Emmans G, Kyriazakis I. Describing and predicting potential growth in the pig. *Animal Science* 2004; **78**: 379–388.

[18] Seber G, Wild C. *Nonlinear regression*, Wiley, New York, 1989.

[19] Burnham K, Anderson DR. *Model Selection and Multimodel Inference: A Practical Information-theoretic Approach*, 2nd Edition, Springer-Verlag, New York, 2002.

[20] Claeskens G, Hjort NL. *Model selection and model averaging*, Cambridge University Press, 2008.

[21] Posada D, Buckey TR. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and bayesian approaches over likelihood ratio tests. *Systematic Biology* 2004; **53**: 793–808.

[22] Moon H *et al.* Model averaging using the kullback information criterion in estimating efective doses for microbial infection and illness. *Risk Analysis* 2005; **25**: 1147–1159.

[23] Lin D, Shkedy Z, Yekutieli D, Amaratunga D, Bijnens L. *Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R*, Springer-Verlag Berlin Heidelberg, 2012.

[24] Faes C *et al.* Model averaging using fractional polynomials to estimate a safe level of exposure. *Risk Analysis* 2007; **27**: 111–123.

[25] Jorgensen M. Fitting animal growth curves. *The New Zealand Statistician* 1981; **16**: 5–15.

[26] Wang X, Wu J, Yang Y. Richards model revisited: Validation by and application to infection dynamics. *Journal of Theoretical Biology* 2012; **313**: 12–19.

[27] Burnham K, Anderson D. Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Researchs* 2004; **33**: 261–304.

[28] Akaike H. *Information theory and an extension of the maximum likelihood principle*, in: B. N. Petrov, F. Csaki (Eds.), Second International Symposium on Information Theory, Akadémiai Kiado, Budapest. 1973, pp. 267–281.

[29] R Core Team. R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2016. `https://www.R-project.org/`

[30] Pinheiro J *et al.* R Core Team, nlme: Linear and Nonlinear Mixed Effects Models, r package version 3.1-128, 2016. `http://CRAN.R-project.org/package=nlme`

[31] Nishiura H. Real-time forecasting of an epidemic using a discrete time stochastic model: a case study of pandemic influenza (h1n1-2009). *BioMedical Engineering OnLine* 2011; **10**: 15.

8.7 Supplementary material

8.7.1 Real time prediction using nonlinear models

A real time prediction is a procedure in which the final size of the epidemic is predicted in an early stage of the outbreak. Let t' be a time point within the outbreak interval and let T be the last time of the outbreak (i.e. for $t > T$ there are no cases , see Figure 8.4 b). The time interval is divided into two periods, the first period from $t = 1$ to $t' \leq T$ is the estimation period. The unknown parameters of the model are estimated using the data within the estimation period and a model based prediction is used to calculate the final size of the epidemic and the turning point. Note that, as shown in the Figure 8.4 a, the final size of the epidemic is a parameter in the model.

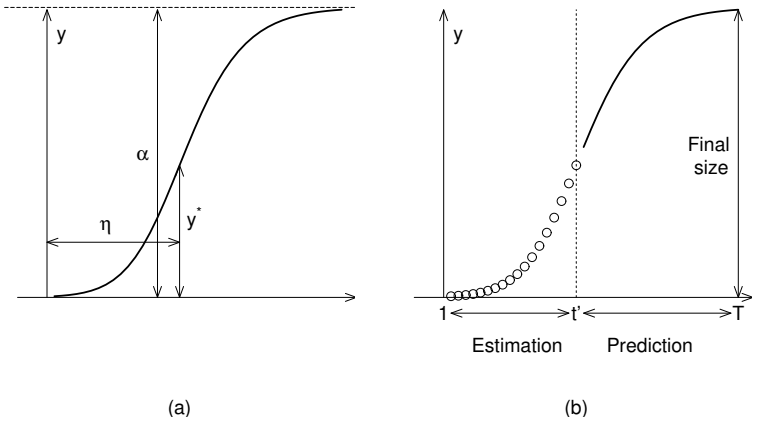


Figure 8.4: Real time prediction using nonlinear models: (a) parameters from Richards model, asymptote α (final size of the outbreak) and η (turning point of an epidemic); (b) real time prediction.

8.7.2 The performance of the Weibull model

Figure 8.3 in the paper reveals an "unusual performance of the Weibull model in two cities: Bucaramanga and Cali". In this section we discuss in more details the performance of the Weibull model. Figure 8.5 shows the observed and predicted number of cases, obtained for the Weibull model when all data are used to estimate the model. Based on these results the models weights, presented in Table 8.3 are calculated. As we mentioned in the results section, for all the cities, the weight of the Weibull model, due to a poor fit to the data, is very small and does not highly influence the parameter estimates and their standard errors. Figure 8.6 presents the real-time prediction in the four cities and reveals that in all cities, the Weibull model has the tendency to over estimate the final size of the epidemic. For Cali and Bucaramanga, when the estimation period is 1-32, as shown in Figure 8.7, the predicted final size is closer to the observed 12380.4 and 3498.2 for Cali and Bucaramanga, respectively (see also in Table 8.4). These estimates are closer to the observed values as can be seen for the final size in Figure 8.6 and the turning point in Figure 8.8.

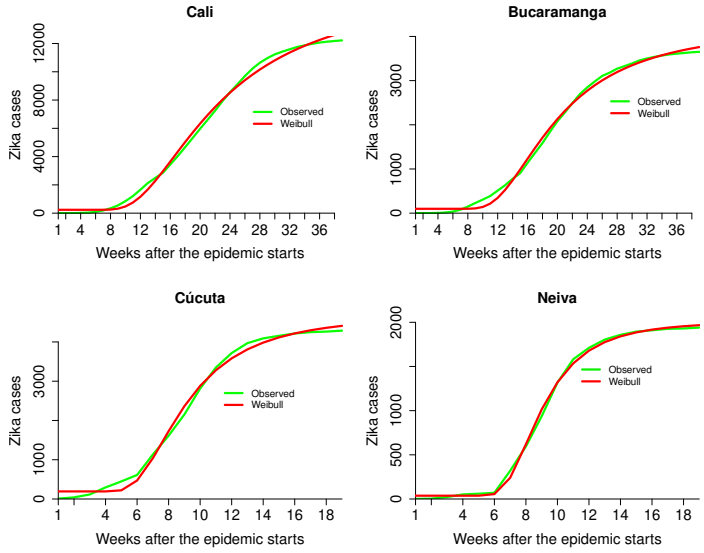


Figure 8.5: Observed and fitted cumulative Zika case counts from the Weibull model in four cities from Colombia. Estimation period 1 to t, where t = T, and T is the maximum number of weeks of the outbreak.

Table 8.4: Parameter estimates for Weibull model in the four cities. EP: estimation period in weeks.

	Cali				Bucaramanga			
EP	α	β	k	η	α	β	k	η
1-22	12029.5	-61.9	-3.1	22.6	4180.2	-1.8	-3.3	22.6
1-23	12720.7	-66.5	-3.0	23.2	3890.3	0.8	-3.3	21.8
1-24	13049.2	-69.1	-3.0	23.5	3667.2	4.0	-3.4	21.2
1-25	13044.4	-69.0	-3.0	23.5	3540.4	6.8	-3.5	20.8
1-26	13159.3	-70.5	-3.0	23.6	3495.8	8.2	-3.5	20.7
1-27	13214.7	-71.3	-3.0	23.7	3449.0	10.1	-3.6	20.6
1-28	13112.0	-69.3	-3.0	23.5	3443.9	10.4	-3.6	20.6
1-29	43296.5	43.2	1.0	38.8	6103.1	54.7	1.8	21.2
1-30	34955.7	61.7	1.2	32.8	5552.7	61.4	1.9	20.1
1-31	29075.9	82.6	1.3	28.6	5218.3	66.7	2.0	19.4
1-32	12380.4	-41.6	-3.1	22.8	3498.2	5.3	-3.5	20.7
1-33	12311.2	-36.9	-3.1	22.7	4778.5	76.4	2.2	18.6
1-34	20456.4	145.6	1.6	22.5	4629.7	80.7	2.3	18.3
1-35	19037.4	165.4	1.8	21.5	4510.9	84.7	2.4	18.1
1-36	17968.0	184.4	1.9	20.8	4412.5	88.5	2.5	17.9
1-37	17129.1	202.7	2.0	20.2	4330.7	92.1	2.6	17.8
1-38	16463.6	220.3	2.1	19.8	4264.0	95.4	2.6	17.7
1-39	15927.9	237.1	2.2	19.5	4206.8	98.5	2.7	17.6
	Cúcuta				Neiva			
EP	α	β	k	η	α	β	k	η
1-12	16158.9	82.2	1.1	16.7	-	-	-	-
1-13	8526.0	110.7	1.7	10.9	2255.7	3.8	8.6	30.0
1-14	6358.3	135.8	2.1	9.3	2152.4	4.1	8.5	31.9
1-15	5506.6	155.0	2.5	8.7	2097.4	4.3	8.4	33.4
1-16	5121.4	168.4	2.8	8.4	2059.9	4.4	8.4	34.8
1-17	4903.8	178.5	3.0	8.3	2037.2	4.5	8.4	35.8
1-18	4762.5	186.9	3.2	8.2	2019.7	4.6	8.4	36.7
1-19	4673.9	193.1	3.3	8.2	2008.6	4.7	8.3	37.4

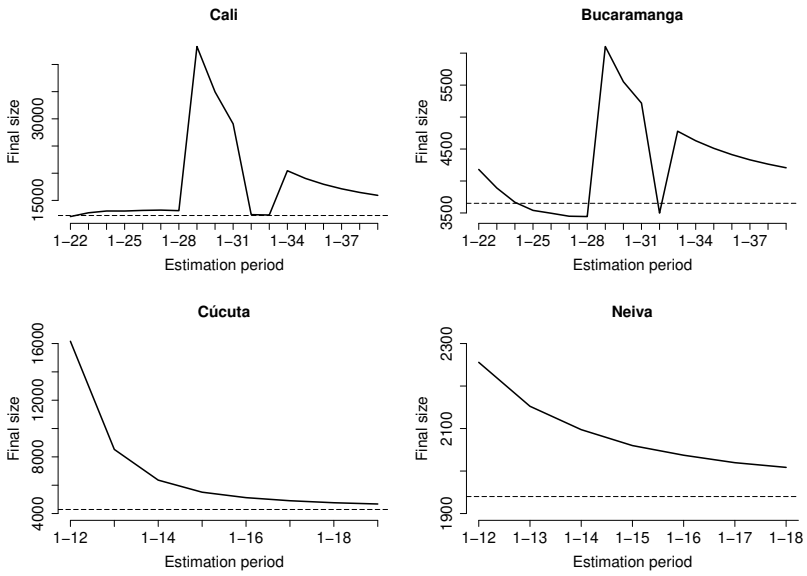


Figure 8.6: Real-time prediction of the final size of the Zika outbreak obtained for the Weibull model in four cities from Colombia. Dashed lines are observed values.

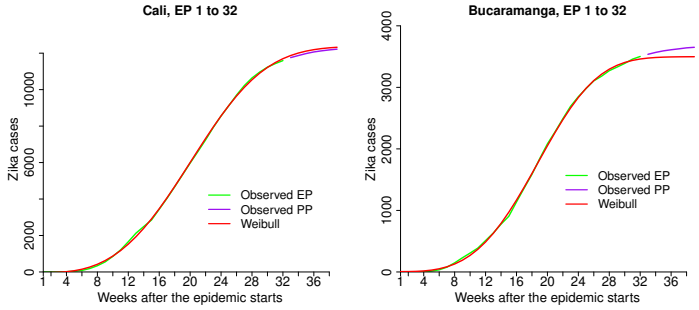


Figure 8.7: Observed and fitted cumulative Zika case counts from the Weibull model in Bucaramanga and Cali. In both cities, the estimation period consist of the first 32 weeks of the outbreak. EP: estimation period; PP: prediction period.

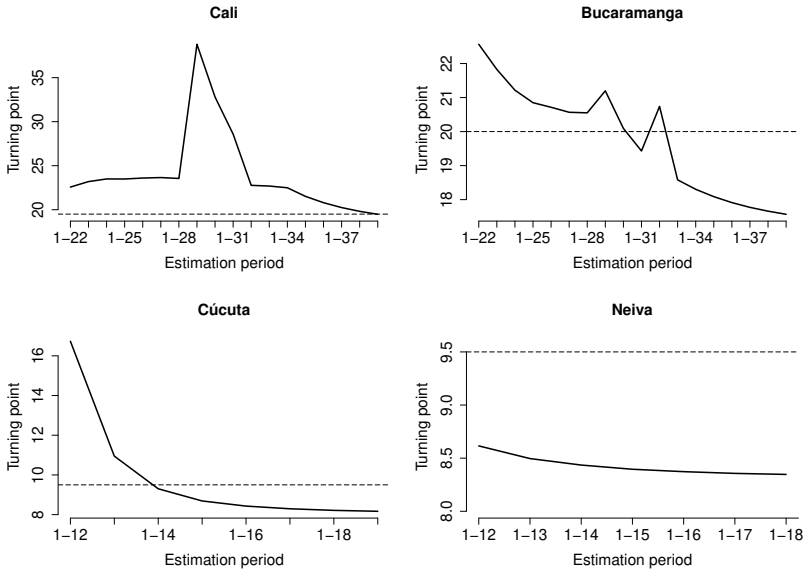


Figure 8.8: Parameter estimates for the turning point for the Zika outbreak obtained for the Weibull model in four cities from Colombia. Dashed lines are observed values

Real time estimation of the turning point and predicted final size (mean and 95% confidence intervals) of the outbreak in four Colombian cities using all nonlinear models and model average for different estimations periods

Table 8.5: Real time estimation of the turning point and predicted final size of the outbreak using all nonlinear model and model average for different estimation periods in the city of Bucaramanga.

Estimated turning point of the outbreak

Period	3P logistic model	Richards model	5p logistic model	Sigmoid Emax	Gompertz model	Weibull model	Model averaging
1-22	18.1 (17.5,18.8)	17.3(16.4,18.2)	21.8(19.6,24)	20.7 (17.5,24)	25.7 (21.9,29.5)	22.6 (20.4,24.7)	17.3 (12.6,21.9)
1-23	18.3 (17.8,18.9)	18 (17.1,19)	20.8 (19.2,22.4)	19.5 (18.1,20.9)	24.1 (21.6,26.6)	21.8 (20.5,23.2)	18 (15,21.1)
1-24	18.4 (18,18.9)	18.3 (17.6,18.9)	19.8 (18.6,21)	18.8 (18.1,19.5)	22.7 (21,24.4)	21.2 (20.4,22.1)	18.3 (16.2,20.4)
1-25	18.5 (18.2,18.8)	18.4 (17.9,18.9)	19 (18.1,20)	18.5 (18.1,18.9)	21.8 (20.5,23)	20.8 (20.3,21.4)	18.4 (16.8,20)
1-26	18.6 (18.3,18.9)	18.6 (18.2,18.9)	18.5 (17.8,19.3)	18.4 (18.1,18.7)	21.1 (20.2,22.1)	20.7 (20.3,21.1)	18.6 (17.4,19.8)
1-27	18.6 (18.4,18.8)	18.6 (18.3,18.8)	18.1 (17.4,18.7)	18.4 (18.1,18.6)	20.6 (19.8,21.4)	20.6 (20.2,20.9)	18.6 (17.6,19.6)
1-28	18.6 (18.4,18.8)	18.7 (18.5,18.9)	17.8 (17.2,18.3)	18.4 (18.1,18.6)	20.2 (19.6,20.9)	20.6 (20.3,20.8)	18.7 (17.9,19.5)
1-29	18.7 (18.5,18.8)	18.7 (18.6,18.9)	17.5 (17.1,18)	18.4 (18.1,18.6)	20 (19.4,20.5)	21.2 (18.5,23.9)	18.7 (18,19.4)
1-30	18.7 (18.5,18.9)	18.8 (18.6,18.9)	17.4 (17,17.7)	18.4 (18.1,18.6)	19.8 (19.3,20.2)	20.1 (18.1,22.1)	18.8 (18.2,19.3)
1-31	18.7 (18.6,18.9)	18.9 (18.7,19)	17.3 (16.9,17.6)	18.3 (18.1,18.6)	19.7 (19.3,20.1)	19.4 (17.9,21)	18.9 (18.4,19.4)
1-32	18.8 (18.6,18.9)	18.9 (18.8,19.1)	17.2 (16.9,17.5)	18.3 (18.1,18.6)	19.6 (19.3,20)	20.7 (20.6,20.9)	18.9 (18.5,19.4)
1-33	18.8 (18.7,19)	19 (18.8,19.1)	17.1 (16.9,17.4)	18.3 (18.1,18.6)	19.5 (19.2,19.8)	18.6 (17.6,19.6)	19 (18.6,19.4)
1-34	18.9 (18.7,19)	19 (18.9,19.1)	17.1 (16.8,17.4)	18.3 (18.1,18.6)	19.5 (19.2,19.8)	18.3 (17.5,19.1)	19 (18.7,19.4)
1-35	18.9 (18.8,19.1)	19 (18.9,19.1)	17.1 (16.8,17.3)	18.3 (18.1,18.6)	19.4 (19.2,19.7)	18.1 (17.4,18.8)	19 (18.7,19.4)
1-36	18.9 (18.8,19.1)	19.1 (19,19.2)	17 (16.8,17.3)	18.3 (18.1,18.5)	19.4 (19.2,19.6)	17.9 (17.3,18.5)	19.1 (18.8,19.4)
1-37	19 (18.8,19.1)	19.1 (19,19.2)	17 (16.8,17.2)	18.3 (18.1,18.5)	19.4 (19.1,19.6)	17.8 (17.2,18.3)	19.1 (18.8,19.3)
1-38	19 (18.9,19.1)	19.1 (19,19.2)	17 (16.8,17.2)	18.3 (18.1,18.5)	19.3 (19.1,19.5)	17.7 (17.1,18.2)	19.1 (18.8,19.3)
1-39	19 (18.9,19.1)	19.1 (19,19.2)	17 (16.8,17.2)	18.3 (18.1,18.5)	19.3 (19.1,19.5)	17.6 (17.1,18)	19.1 (18.8,19.3)

Predicted final size of the outbreak

Period	3P logistic model	Richards model	5p logistic model	Sigmoid Emax	Gompertz model	Weibull model	Model averaging
1-22	3281 (3025,3537)	2846 (2418,3275)	6760 (5244,8276)	5835 (2951,8719)	6738 (4785,8691)	4180 (3345,5016)	2854 (451,5257)
1-23	3371 (3166,3575)	3193 (2735,3651)	6046 (5049,7042)	4784 (3503,6065)	5910 (4762,7058)	3890 (3414,4367)	3198 (1781,4615)
1-24	3415 (3256,3574)	3304 (2985,3624)	5416 (4717,6116)	4193 (3537,4850)	5259 (4523,5994)	3667 (3380,3954)	3305 (2398,4212)
1-25	3441 (3317,3566)	3368 (3146,3589)	4954 (4436,5473)	3906 (3515,4297)	4823 (4309,5336)	3540 (3354,3726)	3368 (2733,4002)
1-26	3470 (3370,3571)	3452 (3277,3627)	4659 (4262,5055)	3798 (3533,4063)	4568 (4188,4947)	3496 (3367,3625)	3452 (2982,3922)
1-27	3477 (3396,3558)	3457 (3334,3579)	4397 (4079,4715)	3687 (3503,3871)	4337 (4040,4633)	3449 (3357,3541)	3457 (3088,3825)
1-28	3492 (3424,3559)	3498 (3399,3597)	4234 (3978,4490)	3652 (3516,3789)	4202 (3966,4439)	3444 (3374,3514)	3498 (3204,3792)
1-29	3502 (3445,3558)	3519 (3441,3597)	4105 (3893,4316)	3625 (3520,3729)	4095 (3901,4289)	6103 (4807,7399)	3519 (3277,3761)
1-30	3514 (3464,3563)	3545 (3479,3611)	4015 (3839,4192)	3616 (3533,3700)	4022 (3860,4184)	5553 (4595,6510)	3545 (3343,3748)
1-31	3533 (3488,3578)	3587 (3524,3650)	3964 (3815,4112)	3630 (3560,3700)	3984 (3848,4121)	5218 (4473,5964)	3587 (3416,3758)
1-32	3550 (3508,3592)	3619 (3561,3678)	3923 (3796,4050)	3642 (3582,3701)	3954 (3837,4071)	3498 (3462,3535)	3619 (3472,3766)
1-33	3565 (3526,3604)	3643 (3590,3696)	3891 (3781,4001)	3651 (3599,3702)	3929 (3828,4031)	4779 (4286,5271)	3643 (3515,3770)
1-34	3579 (3542,3615)	3661 (3613,3709)	3865 (3769,3961)	3659 (3614,3704)	3909 (3820,3998)	4630 (4216,5043)	3661 (3549,3773)
1-35	3591 (3556,3625)	3675 (3632,3718)	3845 (3760,3930)	3666 (3626,3705)	3891 (3812,3971)	4511 (4158,4864)	3675 (3575,3775)
1-36	3601 (3568,3633)	3685 (3646,3724)	3827 (3751,3903)	3670 (3635,3705)	3876 (3805,3947)	4413 (4107,4718)	3685 (3595,3774)
1-37	3609 (3578,3640)	3691 (3656,3726)	3812 (3744,3880)	3674 (3642,3705)	3861 (3797,3926)	4331 (4063,4598)	3691 (3610,3772)
1-38	3617 (3588,3647)	3697 (3665,3729)	3800 (3738,3862)	3677 (3649,3706)	3850 (3791,3908)	4264 (4027,4501)	3697 (3623,3771)
1-39	3624 (3596,3652)	3701 (3672,3729)	3789 (3733,3846)	3680 (3654,3706)	3839 (3785,3892)	4207 (3995,4419)	3701 (3633,3768)

Table 8.6: Real time estimation of the turning point and predicted final size of the outbreak using all nonlinear model and model average for different estimation periods in the city of Cali

Estimated turning point of the outbreak

Period	3P logistic model	Richards model	5p logistic model	Sigmoid Emax	Gompertz model	Weibull model	Model averaging
1-25	19.1 (18.3,19.8)	63.9 (-42.3,170.2)	20.7 (19.7,21.6)	23.3 (20,26.7)	25.6 (23.7,27.5)	23.5 (22.2,24.8)	28.7 (4.4,53.1)
1-26	19.4 (18.7,20.1)	35 (15.4,54.6)	20.6 (19.8,21.4)	22 (20.1,23.8)	25.4 (23.9,27)	23.6 (22.6,24.6)	24 (15.7,32.3)
1-27	19.7 (19.1,20.4)	27.8 (20.5,35)	20.5 (19.9,21.1)	21.1 (20,22.2)	25.2 (24,26.4)	23.7 (22.9,24.4)	23.3 (17.3,29.3)
1-28	20 (19.4,20.5)	24.2 (21,27.3)	20.3 (19.7,20.8)	20.4 (19.7,21.1)	24.7 (23.7,25.7)	23.5 (22.9,24.2)	23.4 (20.5,26.3)
1-29	20.1 (19.6,20.6)	22.2 (20.7,23.8)	19.9 (19.5,20.4)	19.9 (19.4,20.3)	24.1 (23.2,25)	38.8 (29.6,48)	22.5 (19.6,25.4)
1-30	20.2 (19.7,20.6)	21.3 (20.4,22.2)	19.6 (19.1,20)	19.5 (19.2,19.9)	23.5 (22.7,24.3)	32.8 (26,39.6)	21.3 (20.2,22.5)
1-31	20.2 (19.8,20.6)	20.8 (20.3,21.4)	19.2 (18.8,19.6)	19.4 (19,19.8)	22.9 (22.2,23.6)	28.6 (23.6,33.6)	20.8 (19.9,21.7)
1-32	20.3 (19.9,20.6)	20.6 (20.2,20.9)	19 (18.6,19.3)	19.3 (18.9,19.7)	22.4 (21.8,23.1)	22.8 (22.5,23)	22.6 (21.9,23.4)
1-33	20.3 (20,20.6)	20.5 (20.2,20.7)	18.7 (18.4,19.1)	19.3 (18.9,19.7)	22.1 (21.5,22.7)	22.7 (22.5,22.9)	22.5 (21.5,23.5)
1-34	20.3 (20,20.6)	20.4 (20.2,20.6)	18.6 (18.2,18.9)	19.3 (18.9,19.7)	21.8 (21.2,22.3)	22.5 (20.2,24.8)	20.4 (19.7,21.1)
1-35	20.3 (20,20.6)	20.4 (20.2,20.5)	18.4 (18.1,18.7)	19.4 (18.9,19.8)	21.5 (21,22)	21.5 (19.7,23.4)	20.4 (19.7,21)
1-36	20.3 (20.1,20.5)	20.3 (20.2,20.5)	18.3 (18,18.6)	19.4 (19,19.8)	21.4 (20.9,21.8)	20.8 (19.3,22.3)	20.3 (19.8,20.9)
1-37	20.3 (20.1,20.5)	20.3 (20.2,20.5)	18.2 (17.9,18.5)	19.4 (19,19.8)	21.2 (20.8,21.6)	20.2 (19,21.5)	20.3 (19.8,20.9)
1-38	20.3 (20.1,20.5)	20.3 (20.2,20.5)	18.1 (17.9,18.4)	19.4 (19,19.9)	21.1 (20.7,21.4)	19.8 (18.7,20.9)	20.3 (19.8,20.8)
1-39	20.3 (20.1,20.5)	20.3 (20.2,20.4)	18.1 (17.8,18.4)	19.5 (19.1,19.9)	20.9 (20.6,21.3)	19.5 (18.5,20.4)	20.3 (19.9,20.8)

Predicted final size of the outbreak

Period	3P logistic model	Richards model	5p logistic model	Sigmoid Emax	Gompertz model	Weibull model	Model averaging
1-25	10836 (10007,11665)	56912 (-32807,146632)	17507 (15853,19162)	27292 (16648,37936)	19046 (16611,21481)	13044 (11767,14322)	30692 (7137,54246)
1-26	11283 (10519,12046)	29335 (8985,49686)	17414 (16104,18725)	23190 (17177,29202)	18837 (16944,20729)	13159 (12175,14144)	22966 (13130,32802)
1-27	11666 (10968,12363)	21406 (13355,29456)	17217 (16177,18257)	20549 (16817,24282)	18519 (17042,19996)	13215 (12459,13971)	19004 (12957,25051)
1-28	11943 (11325,12561)	17272 (13576,20968)	16808 (15964,17651)	18302 (15886,20718)	17933 (16763,19102)	13112 (12549,13675)	15548 (9578,21518)
1-29	12110 (11579,12642)	14940 (13033,16848)	16238 (15503,16973)	16472 (14817,18127)	17162 (16175,18150)	43296 (31106,55487)	15792 (12994,18591)
1-30	12210 (11757,12663)	13796 (12662,14929)	15669 (15000,16338)	15218 (14026,16410)	16428 (15558,17298)	34956 (25986,43925)	13865 (12617,15113)
1-31	12256 (11871,12640)	13136 (12402,13870)	15126 (14502,15751)	14325 (13432,15217)	15752 (14964,16541)	29076 (22363,35789)	13137 (12147,14128)
1-32	12278 (11950,12606)	12795 (12281,13308)	14670 (14094,15246)	13739 (13051,14426)	15201 (14490,15912)	12380 (12172,12589)	12406 (12070,12742)
1-33	12299 (12015,12583)	12650 (12265,13036)	14319 (13798,14840)	13383 (12841,13925)	14788 (14156,15420)	12311 (12141,12481)	12340 (12038,12642)
1-34	12311 (12063,12559)	12560 (12260,12860)	14028 (13557,14499)	13133 (12694,13571)	14450 (13884,15015)	20456 (17258,23655)	12560 (11848,13272)
1-35	12323 (12103,12542)	12518 (12275,12760)	13798 (13373,14223)	12966 (12605,13328)	14184 (13678,14690)	19037 (16425,21650)	12518 (11880,13156)
1-36	12333 (12138,12529)	12496 (12296,12697)	13612 (13228,13996)	12851 (12548,13154)	13970 (13515,14425)	17968 (15794,20142)	12496 (11922,13071)
1-37	12341 (12165,12517)	12480 (12311,12649)	13456 (13107,13805)	12763 (12504,13022)	13790 (13378,14202)	17129 (15290,18968)	12480 (11959,13000)
1-38	12346 (12187,12506)	12468 (12323,12612)	13326 (13007,13644)	12697 (12473,12921)	13639 (13263,14014)	16464 (14886,18042)	12468 (11993,12942)
1-39	12351 (12205,12497)	12458 (12333,12584)	13216 (12924,13509)	12646 (12449,12842)	13510 (13167,13854)	15928 (14557,17299)	12458 (12023,12894)

Table 8.7: Real time estimation of the turning point and predicted final size of the outbreak using all nonlinear model and model average for different estimation periods in the city of Cúcuta

Estimated turning point of the outbreak

Period	3P logistic model	Richards model	5p logistic model	Sigmoid Emax	Gompertz model	Weibull model	Model averaging
1-12	9 (8.7,9.3)	8.8 (8.2,9.4)	9.1 (8.1,10.2)	9 (8.7,9.4)	10.3 (9,11.7)	16.7 (1.5,31.9)	9 (8.4,9.6)
1-13	9 (8.8,9.2)	8.9 (8.6,9.2)	8.6 (8,9.2)	9 (8.7,9.2)	9.7 (9,10.5)	10.9 (7.3,14.6)	9 (8.6,9.3)
1-14	8.9 (8.8,9.1)	8.9 (8.7,9.1)	8.3 (7.9,8.7)	9 (8.7,9.2)	9.4 (8.9,9.9)	9.3 (7.8,10.8)	8.9 (8.6,9.2)
1-15	8.9 (8.8,9)	8.9 (8.8,9)	8.2 (7.9,8.5)	9 (8.7,9.2)	9.2 (8.9,9.6)	8.7 (7.9,9.5)	8.9 (8.6,9.2)
1-16	8.9 (8.8,9)	8.9 (8.8,9)	8.1 (7.8,8.4)	9 (8.8,9.2)	9.1 (8.8,9.4)	8.4 (7.9,9)	8.9 (8.7,9.2)
1-17	8.9 (8.8,9)	8.9 (8.8,9)	8.1 (7.8,8.3)	9 (8.8,9.2)	9.1 (8.8,9.3)	8.3 (7.9,8.7)	8.9 (8.7,9.2)
1-18	8.9 (8.8,9)	8.9 (8.9,9)	8 (7.8,8.3)	9 (8.8,9.2)	9 (8.8,9.2)	8.2 (7.8,8.6)	8.9 (8.7,9.2)
1-19	8.9 (8.8,9)	8.9 (8.9,9)	8 (7.8,8.3)	9 (8.8,9.2)	9 (8.8,9.2)	8.2 (7.8,8.5)	8.9 (8.7,9.1)

Predicted final size of the outbreak

Period	3P logistic model	Richards model	5p logistic model	Sigmoid Emax	Gompertz model	Weibull model	Model averaging
1-12	4419 (4137,4702)	4103 (3333,4872)	6084 (4607,7560)	4546 (3772,5320)	5842 (4413,7270)	16159 (-3433,35751)	4427 (3660,5194)
1-13	4387 (4219,4556)	4227 (3830,4624)	5303 (4540,6065)	4405 (4053,4757)	5210 (4496,5923)	8526 (3933,13119)	4367 (4000,4733)
1-14	4342 (4229,4456)	4231 (4022,4441)	4877 (4405,5348)	4316 (4126,4506)	4855 (4417,5293)	6358 (4435,8281)	4308 (4039,4578)
1-15	4310 (4226,4394)	4231 (4102,4361)	4649 (4324,4974)	4276 (4156,4396)	4659 (4355,4964)	5507 (4426,6587)	4272 (4041,4502)
1-16	4299 (4234,4365)	4249 (4156,4343)	4537 (4298,4775)	4273 (4187,4358)	4561 (4335,4787)	5121 (4407,5835)	4273 (4101,4445)
1-17	4295 (4243,4348)	4264 (4191,4336)	4471 (4286,4657)	4275 (4210,4340)	4502 (4325,4678)	4904 (4387,5420)	4278 (4147,4409)
1-18	4293 (4248,4337)	4271 (4213,4329)	4428 (4278,4578)	4277 (4224,4329)	4459 (4315,4604)	4762 (4366,5159)	4280 (4173,4387)
1-19	4294 (4256,4332)	4280 (4231,4329)	4403 (4277,4528)	4282 (4238,4326)	4434 (4313,4554)	4674 (4356,4992)	4286 (4199,4372)

Table 8.8: Real time estimation of the turning point and predicted final size of the outbreak using all nonlinear model and model average for different estimation periods at the city of Neiva

Estimated turning point of the outbreak

Period	3P logistic model	Richards model	5p logistic model	Sigmoid Emax	Gompertz model	Weibull model	Model averaging
1-13	8.9 (8.8,9.1)	9 (8.8,9.2)	8.4 (8.2,8.6)	8.8 (8.5,9.1)	9 (8.9,9.2)	8.6 (8.3,9)	8.9 (8.5,9.3)
1-14	9 (8.8,9.1)	9 (8.9,9.2)	8.4 (8.2,8.5)	8.8 (8.5,9.1)	9 (8.9,9.2)	8.5 (8.2,8.7)	8.9 (8.6,9.3)
1-15	9 (8.9,9.1)	9 (8.9,9.2)	8.3 (8.2,8.5)	8.8 (8.5,9.1)	9 (8.9,9.2)	8.4 (8.2,8.6)	8.9 (8.6,9.3)
1-16	9 (8.9,9.1)	9 (8.9,9.1)	8.3 (8.2,8.5)	8.8 (8.5,9)	9 (8.9,9.1)	8.4 (8.2,8.6)	9 (8.6,9.3)
1-17	9 (8.9,9.1)	9 (8.9,9.1)	8.3 (8.2,8.4)	8.7 (8.5,9)	9 (8.9,9.1)	8.4 (8.2,8.5)	9 (8.6,9.3)
1-18	9 (8.9,9.1)	9 (9,9.1)	8.3 (8.2,8.4)	8.7 (8.5,9)	9 (9,9.1)	8.4 (8.2,8.5)	9 (8.7,9.3)
1-19	9.1 (9,9.2)	9 (9,9.1)	8.3 (8.2,8.4)	8.7 (8.5,9)	9 (9,9.1)	8.3 (8.2,8.5)	9 (8.7,9.3)

Predicted final size of the outbreak

Period	3P logistic model	Richards model	5p logistic model	Sigmoid Emax	Gompertz model	Weibull model	Model averaging
1-13	1847 (1786,1909)	1919 (1757,2082)	1999 (1878,2121)	1881 (1773,1989)	1947 (1851,2043)	2256 (1971,2541)	1899 (1743,2055)
1-14	1866 (1819,1912)	1926 (1825,2027)	1971 (1891,2051)	1895 (1825,1965)	1943 (1878,2008)	2152 (1967,2338)	1912 (1804,2020)
1-15	1881 (1843,1920)	1935 (1864,2005)	1960 (1902,2018)	1908 (1857,1959)	1944 (1897,1992)	2097 (1965,2230)	1927 (1850,2005)
1-16	1892 (1859,1925)	1937 (1885,1989)	1952 (1908,1997)	1915 (1875,1954)	1944 (1907,1981)	2060 (1959,2161)	1933 (1876,1990)
1-17	1901 (1872,1930)	1941 (1900,1981)	1950 (1914,1985)	1921 (1889,1953)	1945 (1915,1975)	2037 (1957,2117)	1939 (1895,1983)
1-18	1907 (1881,1933)	1942 (1909,1974)	1947 (1917,1977)	1925 (1898,1952)	1945 (1920,1970)	2020 (1954,2086)	1940 (1905,1976)
1-19	1913 (1889,1937)	1944 (1917,1971)	1947 (1921,1972)	1929 (1906,1952)	1946 (1924,1968)	2009 (1953,2064)	1943 (1914,1973)

Chapter 9

Bayesian Nonlinear Models for Estimation and Prediction for Zika outbreak data

Abstract

Early prediction of a final size of vector-borne diseases is of primary interest since it can provide to health authorities a useful tool to design an efficient policy intervention for the outbreak. The current study describes the usage of hierarchical Bayesian nonlinear models for real time estimation and prediction of both the final size and the turning point of the epidemic. The proposed models are applied to weekly number of Zika cases in the outbreak of 2015/2016 in ten Colombian cities.

Keywords: outbreak modeling; Hamiltonian Monte-Carlo; no-u-turn algorithm.

9.1 Introduction

Mathematical and statistical modeling techniques have been recognized as a vital component to prevention and control of infectious diseases [1]. For worlwide 2009 influenza outbreak, early outbreak investigation proved essential to estimate epidemiological parameters such as the R_0 (basic reproductive number), serial intervals, age specific clinical attack rates, and the rate of secondary infections [2]. Initially, rapid analysis with limited data allowed to generate public health policies, followed by more complex analysis when more reliable data was available, to understand the disease dynamics. One of the modeling techniques correspond to the nonlinear models for the estimation of epidemiological parameters such as the final size and the turning point of an outbreak. Zhou and Yan [3]

analyzed the dynamics of cumulative severe acute respiratory syndrome (SARS) cases in Singapore, Hong Kong, and Beijing using the Richards nonlinear model [4], establishing outbreak key parameters such as the final size of the epidemic and the turning point of the epidemic (the time at which the maximum growth rate is attained), by using the asymptote parameter and the inflection point parameter obtained from the Richards models. Real-time estimation and prediction can be conducted for the epidemiological parameters of interest providing a useful tool for surveillance and policy making in public health. Hsieh et al. [5] used the Richards model to the cumulative SARS case data in Taiwan, making projections of an ongoing outbreak, fitting the model using various intervals around the turning point of the epidemic. Final size and turning point estimates of multi-wave outbreaks can be estimated by the Richards model as shown for the 2003 SARS multiphase outbreak in Toronto [6] and China, Singapore and Canada [7]. Further, [8] applied the Richards model to weekly notifications of Dengue disease case counts in Singapore to identify the turning point of the outbreak and to ascertain the impact of intervention and control measures implemented during the 2005 dengue outbreak in Singapore. One of the main conclusions is that estimating the turning point, with high incidence values of dengue provides to the health authorities a guide to continue the implementation of mass vector control. Hsieh and Chen [9] used the Richards model in order to estimate the turning point and reproduction number to dengue case data from Taiwan, estimating the impact of climatological events for multi-wave dengue outbreaks using a distributed lag time series model including temperature and precipitation. [10] fitted the Richards model to the early influenza case data of 2009 in Mexico and Mexico city considering as a baseline the influenza cases for years 2004-2008, during the same period, estimating the turning point of early influenza infections and the transmissibility of the virus in terms of its basic reproduction number, concluding that there may have been an early epidemic of Influenza A (H1N1) in Mexico City as well as in all of Mexico during February/March 2009. Mostaço et al. [11] established the variability in transmissibility of the 2009 H1N1 pandemic in Canadian communities using the parameters from a Richards model, showing that a considerable geographic variation occurred in transmissibility of the 2009 pandemic outbreaks, precluding the use of transmission projections for large-scale epidemics to the early spread of disease in remote and isolated communities.

Wang et al. [12] made a connection between the Richards model and a Susceptible-Infected-Recovered (SIR) model and used the cumulative infected cases to estimate four parameters: final size, turning point, basic reproduction number and infectious rate, and compared the model with the empirical Richards model fitted to the data. The Richards model has also been employed by Chan et al. [13] to historical data of the Cholera pandemic from different citiesin Europa and North America; Hsieh et al. [14] explored the Richards model for multiwave dengue epidemics in Cuba; Ma et al. [15] estimated the initial epidemic growth rates to cumulative differentiated data; and Hsieh et al. [16] fitted the Richard model for the 2014 Ebola epidemic in Africa. Nonlinear least squares techniques [17] were employed for all previous examples of nonlinear models fitted to cumulative case counts of infectious diseases.

Nonlinear models have been fitted within the Bayesian framework as well [18]. Bolker et al. [19] reviewed Bayesian open source tools available to fit nonlinear models to data

from ecology studies, while the prior selection for Bayesian nonlinear models have been studied by Bornkamp et al. [20]. Klaunberg et al. [21] discussed the choice of prior distributions for the Bayesian nonlinear 4-parameter logistic function applied to calibration of diagnostic curves, while Huson et al. [22] fitted Bayesian logistic nonlinear models to phase I dose-response data. Lin et al. [23] reviewed several nonlinear models such as the 3 and 4 logistic models and the Gompertz model in the context of dose-response studies. Zika virus (genus *flavivirus*, family *flaviviridae*) spread rapidly throughout Central and South-America, affecting near 700,000 persons in 48 countries in 2016, and it was associated with 2,439 cases of a congenital syndrome [24]. Colombia was a country severely affected by the Zika epidemic during the 2015-2016 outbreak [25]. Statistical modeling techniques have been applied to study the epidemiology and transmissibility of Zika in two Colombian cities ([26]); forecasting Zika incidence daily case counts from a Colombian province using nonlinear models (Generalized Richards model and logistic model) [27]; the estimation of the R_0 and the role of sexual transmission in one Colombian city [28]; and the real-time estimation and prediction of zika outbreaks using weekly data and the Richards, the 3-parameter logistic, the 5-parameter logistic, the Sigmoid Emax, the Weibull, and the 4-parameter Gompertz models in four Colombian cities [29].

The analysis presented in this paper is focused on hierarchical Bayesian nonlinear modeling and aim to conduct a real-time estimation and prediction the final size and turning point of the epidemic. The proposed method is applied to outbreak data of the 2015-2016 Zika epidemic in ten Colombian cities with the highest case counts in the outbreak.

The paper is organized as follows. In Section 2, the Zika outbreak data in 10 cities from Colombia are presented. In section 3, the hierarchical Bayesian nonlinear models for cumulative number of Zika cases are formulated, and the results are presented in Section 4. Concluding remarks and given in Section 5.

9.2 The Zika Outbreak Data in 10 cities from Colombia

Ten Colombian cities with the highest cumulative counts up to the 42nd epidemiological week of 2016 were included in the analysis [30]. Table 9.1 shows the summary of the data. Most of the cities started the outbreak between October and December 2015. Cali was the city with the highest case count, followed by Cúcuta, Bucaramanga and Neiva. The lowest case count was for Girón. Bucaramanga, Cúcuta, Floridablanca and Neiva presented the highest incidence of the cities. Bucaramanga, Floridablanca and Girón are closed cities, distant no more than 20 kilometers. Figure 9.1 displays the Zika incidence case counts in the ten cities. Note that the time scale in Figure 9.1 is the epidemiological week after the first January, 2015. Figure 9.1 reveals differences between the cities. For example, Bucaramanga, Cúcuta, Cali and Neiva show a symmetric curve, while Barranquilla, Ibagué, Girón, Palmira and Villavicencio display a sligth skewness to the left, with Palmira displaying several peaks throughout the outbreak period. While Figure 9.1 shows the incident cases, Figure 9.2 presents the Zika cumulative case counts per city.

Table 9.1: Summary information for the Zika outbreak data in ten Colombian cities, 2015-2016

City	Outbreak		Final	Turning	Population	Incidence	Altitude
	start (Date)	Length (Weeks)	size	point (Week)	size	(cases per 100000)	(m)
Barranquilla	18 Oct 15	30	1796	14	1,214,253	14.8	18
Bucaramanga	27 Dec 15	38	3655	17	521,446	70.1	959
Cali	20 Dec 15	38	12220	22	2,333,203	52.4	1018
Cúcuta	08 Nov 15	20	4286	12	628,107	68.2	320
Floridablanca	08 Jan 16	37	1736	17	255,833	67.9	925
Girón	17 Jan 16	35	956	13	161,451	59.2	777
Ibagué	15 Nov 15	37	1229	14	523,057	23.5	1285
Neiva	15 Nov 15	20	1941	11	322,445	60.2	442
Palmira	10 Jan 16	35	1408	9	244,385	57.6	966
Villavicencio	03 Jan 16	36	1605	13	460,732	34.8	437

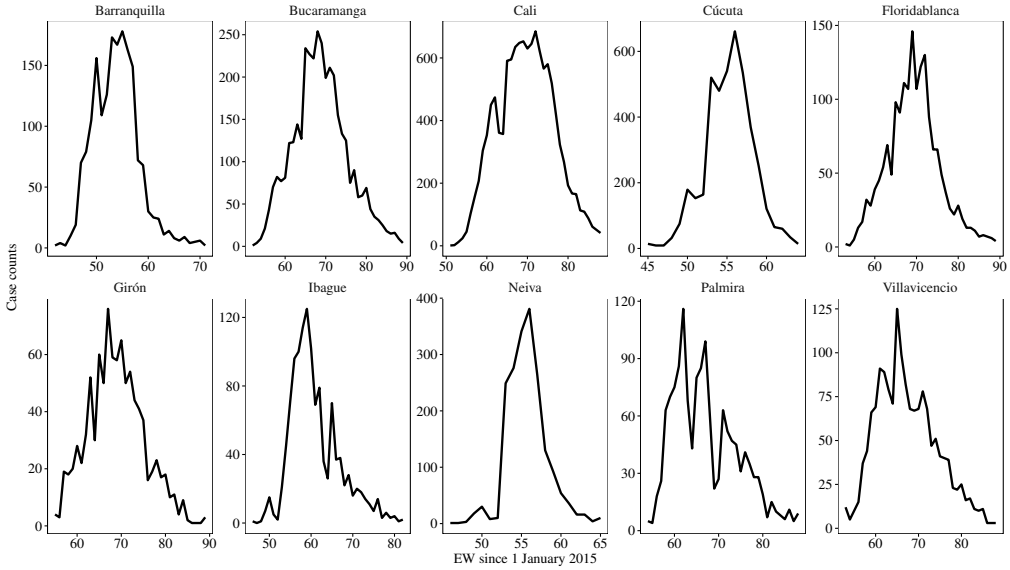


Figure 9.1: Weekly number of Zika cases in ten Colombian cities during the outbreak in 2015-2016.

We appreciate the characteristic sigmoidal form of the single cumulative outbreak data. In Ibagué and Palmira, the sigmoidal patterns is slightly distorted, which reflects the multi-peak behavior shown in the incident plots.

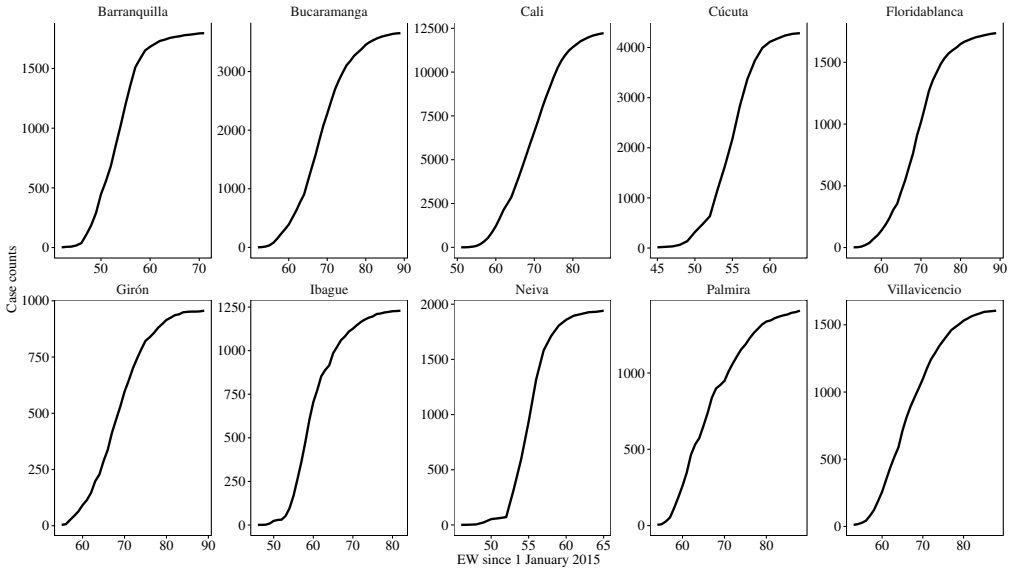


Figure 9.2: Cumulative number of Zika cases in ten cities from Colombia in the outbreak of 2015-2016.

9.3 Hierarchical Bayesian Nonlinear Models for Cumulative Counts

9.3.1 Model formulation

Let Y_t be the cumulative number of Zika cases in the epidemiological week t, $(t = 1, \ldots, T)$. A natural approach to model the count data is to specify a Poisson or Negative Binomial (NB) likelihood for the counts. Since overdispersion is expected the NB that account for overdispersion is chosen. Specifically we assume

$$\begin{aligned} g(y|\mu,\kappa) &= \binom{\kappa+y-1}{y}\left(\frac{\mu}{\mu+\kappa}\right)^{y}\left(\frac{\kappa}{\mu+\kappa}\right)^{\kappa} \\ E(Y) &= \mu(t,\theta) \\ Var(Y) &= \mu+\frac{\mu^2}{\kappa}. \end{aligned}$$

Here, κ corresponds to a dispersion parameter accounting for the expected increasing variance for the cumulative counts. In addition, a Normal likelihood, is specified for the Y_t, that is

$$\begin{aligned} g(y|\mu,\sigma) &= \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2\right) \\ E(Y) &= \mu(t,\theta) \\ Var(Y) &= \sigma^2 \end{aligned}$$

Several models, shown in Table 9.2, were considered for the mean structure of $\mu(t,\theta)$

Table 9.2: Mean structure for the nonlinear models fitted to the Zika outbreak data

Model	$\mu(t,\theta)$	Epidemic shape
Richards	$\frac{\alpha}{[1+\delta\exp(-\gamma(t-\eta))]^{\frac{1}{\delta}}}$	Asymmetric
3P logistic	$\frac{\alpha}{1+\exp(-\gamma(t-\eta))}$	Symmetric
4P logistic	$\beta+\frac{\alpha-\beta}{(1+\exp((\eta-t)/\gamma))}$	Symmetric
5P logistic	$\beta+\frac{(\alpha-\beta)}{\left[1+(2^{(1/g)}-1)(\frac{t}{\eta})^{-k}\right]^{g}}$	Asymmetric
4P Gompertz	$\beta+(\alpha-\beta)\exp(-\exp(-\gamma(t-\eta)))$	Asymmetric

For the Richards model [4], we used the parameterization shown by [31], where α is the horizontal asymptote when $t \to \infty$, γ/δ is the growth rate [12], η corresponds to the

inflection point at t (the turning point), and δ is a parameter with a one-to-one relationship with the basic reproductive number R_0 [12].
For the three-parameter logistic (3PL) model [23] , γ is a scale parameter, α is the horizontal asymptote as $t \to \infty$. The parameter η is the time for which the cumulative number of cases is half of the final size, i.e, $\alpha/2$ which is the inflection point of the curve.
For the four-parameter logistic (4PL) model [23], the parameter β is the horizontal asymptote as $t \to \infty$ and α is the horizontal asymptote as $t \to -\infty$, η is the time for which the outbreak course is in its inflection point (the number of Zika cases in the midway between the asymptotes), and $\gamma > 0$ is a scale parameter on the t-axis [32].
For the five-parameter logistic (5PL) model, we used the parameterization from [33]. The 5PL model presents horizontal lower and upper asymptotes β and α, respectively; η is the inflection point; g is the parameter introduced for asymmetry; and k is a shape parameter which determines the rate at which the outbreak curve rises to the upper asymptote. Note that both the k and g jointly determine the rate at which the number of Zika cases reaches the two asymptotes α and β.
The four-parameter Gompertz model (4P Gompertz) [23] includes the asymptote at $t \to \infty$, α, the asymptote at $t \to$ -∞), β, the scale parameter γ, and the inflection point η.
Both Negative Binomial and Normal likelihoods were specified for the cumulative number of cases. Prior distributions for the unknown parameters in the different models are given in Table 9.3. The following consideration were made for the specification of

Table 9.3: Prior distributions for the unknown parameters in the mean structure of the Negative binomial and Normal models.

	α	η	γ	δ	g	k	β
Richards	U(0, 5×10^5)	U(5, 60)	U(0, 3)	U(-3, 3)			
3PL	U(0, 5×10^5)	U(5, 60)	U(0, 3)				
4PL	U(0, 5×10^5)	U(5, 60)	U(0, 10)				N(0, 100)
5PL	U(0, 5×10^5)	U(5, 60)			U(0, 5)	U(0, 15)	N(0, 100)
4PG	U(0, 5×10^5)	U(5, 60)	U(0, 3)				N(0, 100)

U: Uniform distribution; N: Normal distribution.

the prior distributions: The prior for η was restricted between 5 and 60 weeks since we do not expect to have inflection points (turning points) less than 5 or more than 60 epidemiological weeks.The parameter space of γ is restricted to be non negative. For α the upper bond of 50000 reflects the maximum expected number of cases.
We assigned a Half-Cauchy(0,25) [34] prior distribution to the standard deviation σ in the Normal models, and the dispersion parameter κ in the Negative Binomial models.

9.3.2 Model averaging for Bayesian parameters from nonlinear models

Besides estimating the final size and turning points nonlinear parameters for every model, we were also interested in testing the parameters combination of several models using stacking as a model average procedure. Stacking average point estimates from multiple models [35]. We calculated weighted parameters for the final size (α) and turning point

(η) of the epidemic, using the mean posterior predicted Zika cumulative cases $\hat{Y}_{i,t,j}$ at estimation period 1 to t per city i and nonlinear model j, for the Normal and Negative Binomial likelihoods, and the observed Zika cumulative cases $Y_{i,t}$.
The procedure included: first, obtaining the mean posterior predictive distribution of the Zika cumulative cases $\hat{Y}_{i,t,j}$. Second, we calculated the weights $w_{i,t,j}$ per city i at estimation period 1 to t for the nonlinear models j following [36], by minimizing $||Y_{i,t} - \sum_j w_{i,t,j} \cdot \hat{Y}_{i,t,j}||_2$ using a quadratic approximation with the restriction $\sum_j w_{i,t,j} = 1$. The numerical approximation was implemented using the L-BFGS algorithm of the `optimizing` function from the `rstan` package [37]. The final size and turning point weighted parameters ($\alpha_{W_{i,t}}$ and $\eta_{W_{i,t}}$ respectively) at estimation period 1 to t per city i were calculated as

$$\alpha_{W_{i,t}} = \sum_j w_{i,t,j} \cdot \alpha_{i,t,j}$$
$$\eta_{W_{i,t}} = \sum_j w_{i,t,j} \cdot \eta_{i,t,j}$$

9.3.3 Comparing the final size and turning point from the nonlinear models at several time points with the observed final size and turning point

We calculated in each iteration the absolute percentage error (APE) as the absolute difference between the observed final size and the turning point with the parameters representing the final size (α) and the turning point (η), obtained from the nonlinear models j estimated per city i at estimation periods 1 to t, where $t = 10, \cdots, T-1$ (T is the outbreak length in epidemiological weeks).
The final size $\text{APE}_{\alpha_{i,t,j}}$ of the epidemic was calculated per city i, estimation period 1 to t and nonlinear model j using the equation,

$$\text{APE}_{\alpha_{i,t,j}} = |\alpha_{i,t,j} - K_i| \cdot 100$$

where K_i is the observed maximum Zika cumulative case count per city i. For the calculation of the turning point $\text{APE}_{\eta_{i,t,j}}$ of the epidemic we used,

$$\text{APE}_{\eta_{i,t,j}} = |\eta_{i,t,j} - E_i| \cdot 100$$

where E_i is the observed epidemiological week where the epidemic presented the maximum number of incident cases per city i.
We present the results for the final size estimation and prediction, and turning point estimation using plots of the evolution of the APEs in the estimation periods 1 to $t, \cdots, T-1$, per city i and nonlinear model j, making general remarks from the observed trends. However, interpreting the plot for model selection with the best approximation to the observed values of final size and turning point of the epidemic is not straightforward, therefore, we define a threshold for the APEs of the final size and turning point, and report the week during the outbreak period where the APEs reach the threshold.

For the final size APEs, we fix the threshold at 5%, representing an error of $\pm$ 600 cases in Cali (12220 maximum Zika cumulative case count), or $\pm$ 95 cases in Girón (956 maximum Zika cumulative case count). For the turning point APEs, we fix the threshold at 10 %, representing an error of $\pm$ 2.2 weeks around the observed turning point in Cali, and $\pm$ 0.9 weeks around the observed turning point in Palmira.

9.4 Application to the Data

We fitted the Bayesian nonlinear models for the entire outbreak length (OL) per city (subsection 9.4.1 and subsection 9.4.2), and then, for estimation periods 1 to t, where $t = 10, \cdots$,T-1 (T is the OL per city) (subsection 9.4.3). For the Bayesian inferences, we used the `R` interface to Stan [38] `rstan` [37] . The models were fitted with the no-U-turn sampler [39], and 2000 iterations in total, 1000 iterations of warmup, 1 iteration of thinning and four chains. We compared the model fit using the widely applicable information criterion (WAIC) [40] using the R package `loo` [41].

9.4.1 Model estimation for the complete outbreak length

Table 9.4 shows the WAIC for the models fitted to the entire OL per city. For the Bayesian nonlinear models with Normal likelihood, the 5PL model presented the lowest WAIC for Barranquilla, Cali and Ibagué. The 4PL model showed the lowest WAIC for Bucaramanga, Floridablanca and Girón; and the Richards model for Cúcuta, Neiva, Palmira and Villavicencio.
Although the WAIC for the Normal and Negative Binomial likelihood are not directly comparable, in general the information criteria for the Negative Binomial nonlinear models are higher than the Normal models, and the Richards models are associated to small WAIC values for most of the cities, followed by the 5PL model.

Table 9.4: WAIC for Bayesian nonlinear models with Normal and Negative Binomial likelihood per city, fitted to the entire outbreak length (OL)

		Richards		Logistic 3P		Logistic 4P		Logistic 5P		Gompertz	
City	OL	WAIC	p_{WAIC}	WAIC	p_{WAIC}	WAIC	p_{WAIC}	WAIC	p_{WAIC}	WAIC	p_{WAIC}
						Normal					
Barranquilla	30	281.1	5.4	281.7	4.2	272.0	5.1	**262.5**	6.2	308.3	4.7
Bucaramanga	38	352.2	4.0	386.4	2.7	**346.5**	4.5	346.7	5.8	413.0	5.0
Cali	38	482.8	4.0	497.9	2.6	458.0	4.1	**433.5**	5.4	514.5	4.3
Cúcuta	20	**204.9**	4.7	207.5	4.3	209.1	4.6	207.1	6.7	247.5	4.4
Floridablanca	37	301.0	3.8	301.6	2.8	**292.1**	4.2	293.0	5.2	368.8	4.7
Girón	35	230.2	4.3	265.4	3.0	**216.9**	4.0	226.4	5.3	281.6	5.3
Ibagué	37	320.1	4.5	379.8	4.0	369.4	4.4	**287.7**	5.6	324.4	4.6
Neiva	20	**188.1**	6.1	196.4	5.7	195.9	5.0	188.8	7.3	193.3	5.6
Palmira	35	**300.7**	4.0	377.1	3.2	330.3	6.0	308.3	4.6	314.8	5.3
Villavicencio	36	**269.5**	4.8	366.3	3.1	317.6	5.7	281.7	5.4	269.6	4.6
						Negative binomial					
Barranquilla	30	**287.0**	5.1	355.4	5.8	356.3	23.4	290.0	7.3	291.8	4.0
Bucaramanga	38	442.6	8.4	548.1	9.6	420.7	4.4	**366.0**	3.3	432.1	4.9
Cali	38	**472.9**	3.3	629.0	6.4	603.8	6.5	480.2	4.2	494.5	5.2
Cúcuta	20	236.8	10.2	237.5	10.6	235.0	10.8	**221.7**	9.4	243.9	3.6
Floridablanca	37	357.8	7.2	399.6	6.0	326.9	2.2	**311.5**	1.5	376.3	3.6
Girón	35	299.3	3.8	334.1	4.5	**283.0**	1.2	284.5	1.5	307.9	3.7
Ibagué	37	**343.2**	6.7	374.2	4.0	369.7	4.6	347.6	11.1	361.0	12.1
Neiva	20	**235.4**	8.0	237.0	8.9	238.4	9.4	243.2	7.8	246.7	5.0
Palmira	35	321.1	4.4	436.5	5.7	408.5	9.3	**306.6**	1.9	347.8	6.7
Villavicencio	36	313.8	1.9	418.2	4.7	385.0	10.2	311.0	1.4	**309.8**	1.2

9.4.2 Final size and turning point parameters from the nonlinear models for the complete outbreak length

Figure 9.3 to Figure 9.6 show the mean and 95% credible intervals obtained from the Bayesian nonlinear models for the complete outbreak length, of the final size and turning point of the epidemic of the Normal and Negative Binomial likelihoods. For the final size parameter of the Normal likelihood models (Figure 9.3), the CI of the Richards model included the observed final size for the cities of Barranquilla, Bucaramanga, Cúcuta, Floridablanca, Ibagué and Neiva. In general, the Gompertz model overestimated the final size in most of the cities, while 3PL models tend to underestimate the final size. The weighted parameter for the final size include the observed final size only in four of the cities (Barranquilla, Bucaramanga, Floridablanca and Neiva).

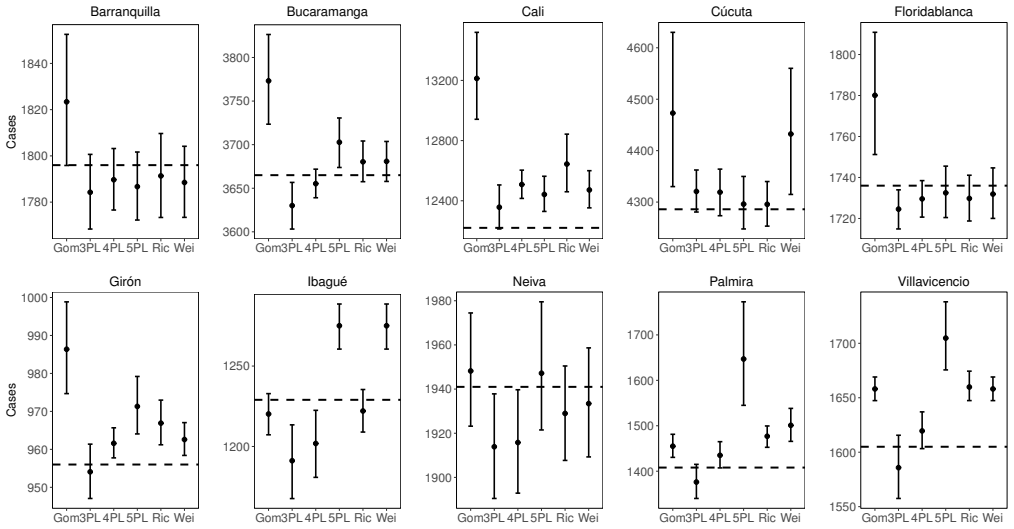


Figure 9.3: Posterior mean and 95% credible intervals for the α parameter (estimating the final size of the epidemic) obtained from the Bayesian nonlinear models with Normal likelihood. Dashed line is the observed final size.

For the turning point parameter of the normal likelihood (Figure 9.4), in general, the nonlinear models underestimated the observed turning point in the cities of Barranquilla, Cali, Cúcuta, Floridablanca and Neiva, while for three cities (Girón, Ibagué and Villavicencio) the nonlinear models rather over or underestimated the observed turning point. In Bucaramanga and Palmira the 95% CI from the 3PL and Richards models respectively included the observed turning point. The turning point weighted parameters were not important in any of the cities.

For the final size parameter obtained from the Negative Binomial likelihood (Figure 9.5), the Richards model included the observed final size in three cities (Barranquilla, Cúcuta and Neiva), while the 3PL and 4PL models underestimated the observed final size in fice cities (Barranquilla, Bucaramanga, Cali, Ibagué and Villavicencio), and the

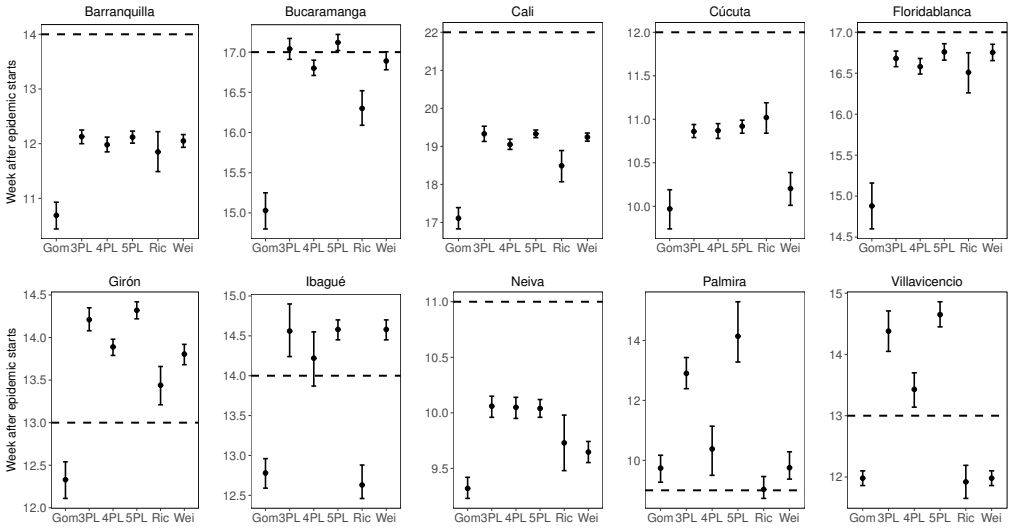


Figure 9.4: Posterior mean and 95% credible intervals for the η parameter (estimating the turning point of the epidemic) obtained from the Bayesian nonlinear models with Normal likelihood. Dashed line is the observed turning point.

Gompertz model did not overestimate the final size of the epidemic in two cities (Ibagué and Palmira). The final size weighted parameters included the observed final size in five cities (Barranquilla, Cúcuta, Floridablanca, Girón and Neiva). In general, the 5PL model overestimated the observed final size in almost all the cities.

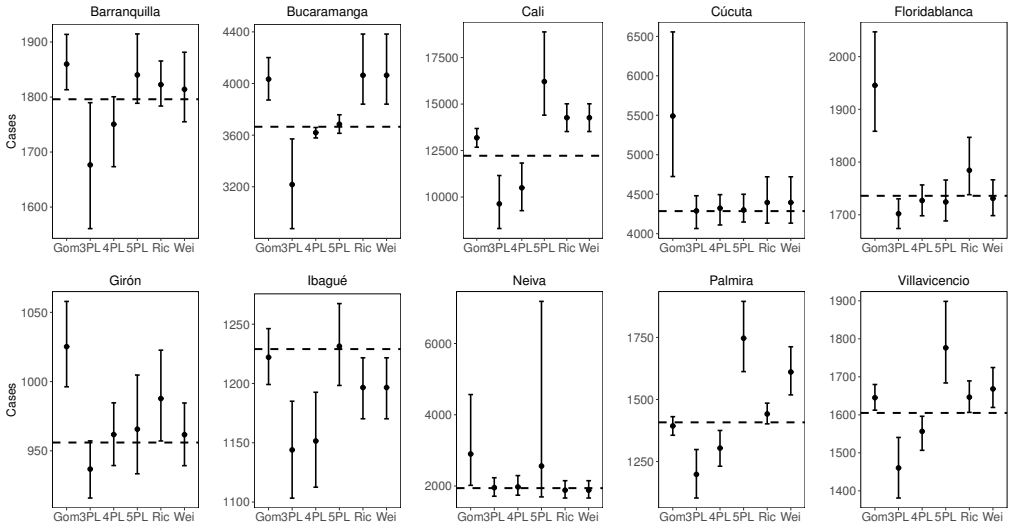


Figure 9.5: Posterior mean and 95% credible intervals for the α parameter (estimating the final size of the epidemic) obtained from the Bayesian nonlinear models with Negative Binomial likelihood. Dashed line is the observed final size.

For the turning point of the nonlinear models with Negative Binomial likelihood (Figure 9.6), all models underestimated the observed turning point in five cities (Barranquilla, Bucaramanga, Cali (with the exception of the 5PL model), Cúcuta, and Floridablanca), while most of the models overestimated the turning point in two cities (Palmira and Girón). The 5PL model included the observed turning point in two cities (Bucaramanga and Neiva), the Richards model in one city (Girón) and the 3PL in one city (Villavicencio).

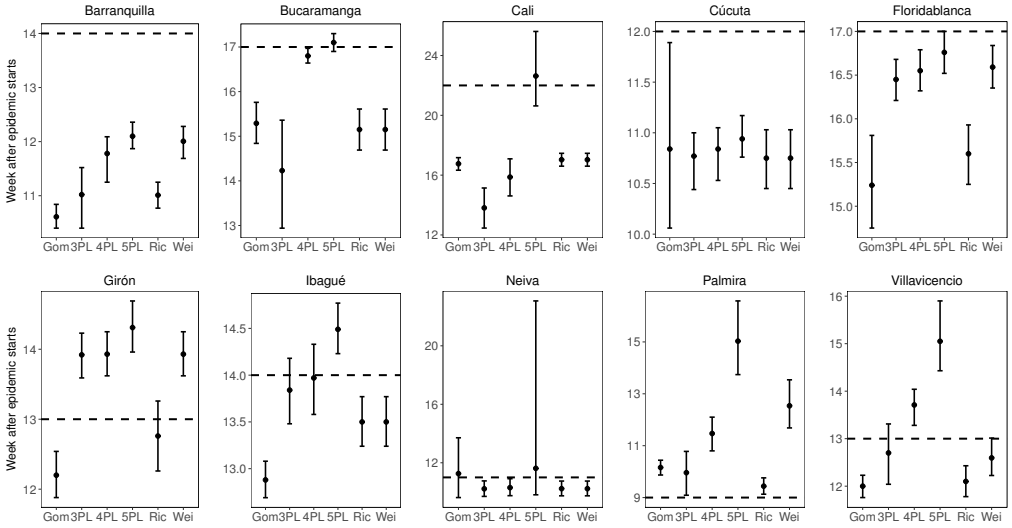


Figure 9.6: Posterior mean and 95% credible intervals for the η parameter (estimating the turning point of the epidemic) obtained from the Bayesian nonlinear models with Negative Binomial likelihood. Dashed line is the observed turning point.

9.4.3 Real time estimation and prediction for the final size and turning point of the outbreak

In this section, we present the results of the real-time estimation of the turning point, and estimation and prediction of the final size of the Zika epidemic in the ten Colombian cities, obtained from the Bayesian nonlinear models, for the Normal and the Negative Binomial models.

Estimation and prediction of the final size of Zika outbreak

Table 9.5 presents the week during the outbreak period where the median of the final size APEs are less than or equal to 5%, per city and model, for the Normal and Negative Binomial likelihood, summarizing the results as follows:

- The final size parameter of the 3PL model for Barranquilla reached the APE 5% threshold by the 55% of the OL, and for Floridablanca by the 50% of the OL.
- In Cúcuta, the final size parameter APEs of the 3PL and Richards models were less than 5% by the 70% of the OL.
- The final size parameter APE of the 4PL model for Bucaramanga was less than or equal to 5% by the 50% of the OL, and for Cali by the 60% of the OL.
- In Palmira, the final size parameter APEs of the Weighted α was less than or equal to 5% by the 50% of the OL, and for Ibagué by the 40% of the OL.
- In Girón, the final size parameter APEs from the 4PL model and the weighted α were less than or equal to 5% by the 48% of the OL while in Neiva, the Richards models and the weighted α by the 65% of the OL.
- In Villavicencio, the final size parameters APE from the 4P Gompertz model reached the 5% threshold by the 45% of the OL.

Results for the final size APEs of the Negative Binomial models are not presented, because the APEs were higher than the Normal models, making them less appropiate for real-time estimation and prediction of the final size of the epidemic. The widely studied Richards model for estimation and prediction of epidemiological parameters did not perform worst than the other models, but the 3PL and 4PL models started making predictions close to the observed values faster than the other models in the outbreak period. Furthermore, we found that the APEs for the weighted α make the prediction for final size in three cities, showing the possibility to make a model 'averaged' final size prediction, instead of a single model prediction.
Figure 9.7 shows the evolution of the $\text{APE}_{\alpha_{i,t,j}}$ for the final size of the epidemic for the Bayesian nonlinear Normal models and the weighted $\text{APE}_{W_{i,t}}$, per city i and estimation periods 1 to $t, \cdots, T-1$.

Table 9.5: Week after the epidemic starts where the absolute percentage error (APE) between the observed final size of the epidemic and the α parameter is less than or equal to 5%, for the nonlinear models

		Normal likelihood						Negative binomial likelihood					
City	OL	Gom	3PL	4PL	5PL	Ric	W	Gom	3PL	4PL	5PL	Ric	W
Barranquilla	30	23	16	19	19	20	19	27	-	27	27	23	23
Bucaramanga	38	33	24	20	26	24	21	-	-	27	26	-	25
Cali	38	-	26	23	31	36	31	26	-	-	-	-	19
Cúcuta	20	20	14	15	15	14	20	-	16	16	17	19	12
Floridablanca	37	32	19	23	27	24	21	-	30	21	27	34	25
Girón	35	31	21	17	21	22	17	-	28	21	26	32	19
Ibagué	37	27	33	31	22	26	15	25	-	-	28	32	16
Neiva	20	14	14	14	14	13	13	-	20	20	-	-	15
Palmira	35	23	30	24	-	21	17	29	-	-	-	25	27
Villavicencio	36	16	29	24	-	19	31	21	-	33	-	23	13

OL: outbreak length.

In conclusion, the final size APEs reached the threshold faster for the 3PL (Barranquilla, Floridablanca, Cucuta), Richards (Cucuta, Neiva), 4PL (Bucaramanga, Cali, Giron) and 4P Gompertz (Villavicencio) models, while the weighted α for Palmira and Ibagué. In addition, the final size APEs reached the threshold by the 40-45% (Ibagué and Villavicencio), 50% - 55% (Barranquilla, Floridablanca, Bucaramanga, Palmira and Girón), 60% - 65% (Cali and Neiva) and 70% (Cúcuta) of the OL.

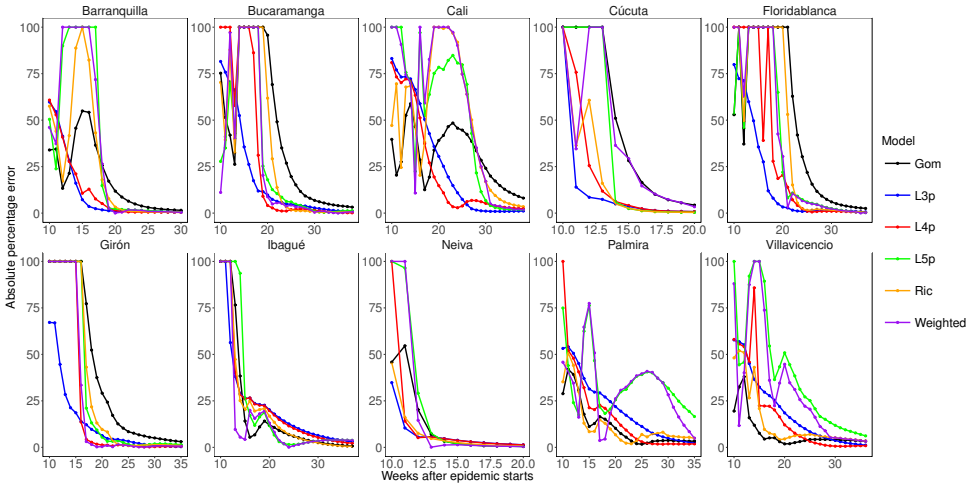


Figure 9.7: Absolute percentage errors (APEs) between the final size parameter (α) and the observed final size of the epidemic per city, from Bayesian nonlinear models with Normal likelihood fitted at estimation periods 1 to t, where $10 \leq t \leq$ T-1, and T is the outbreak length

Estimation of the turning point of Zika outbreak:

Table 9.6 presents the week during the outbreak period where the median of the final size APEs were less than or equal to 5%, per city and model, for the Normal and Negative Binomial likelihood, summarizing the results as follows,

- In Barranquilla, the turning point APE of the Richards model was less than or equal to 10% by the 46% of the OL.
- The 3PL model reached the 10% APE threshold for the turning point parameter in Bucaramanga by the 45% of the OL; in Floridablanca by the 46% of the OL; and Girón by the 37% of the OL.
- For Cali, the weighted η reached the 10% threshold by the 47% of the OL; for Cúcuta by the 55% of the OL; for Neiva by the 50% of the OL; in Palmira by the 31% of the OL; and Villavicencio by the 39% of the OL.

Table 9.6: Week during the outbreak period where the absolute percentage error (APE) between the observed turning point of the epidemic and the η parameter is $\leq 10\%$, for the Bayesian nonlinear models with Normal and Negative Binomial likelihood

		Normal likelihood						Negative binomial likelihood					
City	OL	Gom	3PL	4PL	5PL	Ric	W	Gom	3PL	4PL	5PL	Ric	W
Barranquilla	30	15	-	16	-	14	18	-	-	-	21	-	-
Bucaramanga	38	22	17	19	20	21	19	18	-	24	25	26	11
Cali	38	22	-	-	27	19	18	-	-	-	36	-	-
Cúcuta	20	13	12	12	14	12	11	15	-	17	17	14	15
Floridablanca	37	22	17	18	20	21	21	23	28	18	24	25	24
Girón	35	19	13	17	18	18	17	22	18	19	23	22	19
Ibagué	37	14	15	14	16	14	14	22	14	15	19	16	15
Neiva	20	11	12	12	12	11	10	18	12	13	19	14	14
Palmira	35	14	13	13	-	14	11	14	16	14	-	14	10
Villavicencio	36	14	18	15	-	15	11	23	24	20	-	21	30

OL: outbreak length.

- In Ibagué, the turning point parameters from the 4P Gompertz, 4PL and Richards models, and the weighted η reached the 10% APE threshold by the 38% of the OL.

As observed for the final size estimation, the estimated turning point of the epidemic from the Negative Binomial models presented APEs reaching the threshold only close to the end of the OL, precluding the use of this likelihood for turning point estimation.
Figure 9.8 shows the evolution along the outbreak period of the APEs between the estimated and observed turning point of the epidemic for the Bayesian nonlinear models per city. The plots complements the results for the final size estimation.
Remarkably, in Cúcuta, Palmira and Villavicencio the turning point weighted η reached the APE 10% threshold by the estimation periods 1 to t, where $t = 10, 11$. They reflected the good fit of the nonlinear models in early stages of the outbreak, where the parameters change to fit the data, in spite of the difference between the turning point estimates and the observed turning point.
In summary, the turning point parameters reached the APE 10% threshold faster for the Richards (Barranquilla, ibague), 3PL (Bucaramanga, Floridablanca, Girón, Ibagué), 4PL (Ibagué) and 4P Gompertz (Ibagué) models, while the weighted η for the Cali, Cúcuta, Neiva, Palmira and Villavicencio; and by the 31% (Palmira), 37-39% (Girón, Ibagué and Villavicencio), 45-47% (Barranquilla, Bucaramanga, Floridablanca, Cali) and 50-55% (Cucuta, Neiva) of the OL.

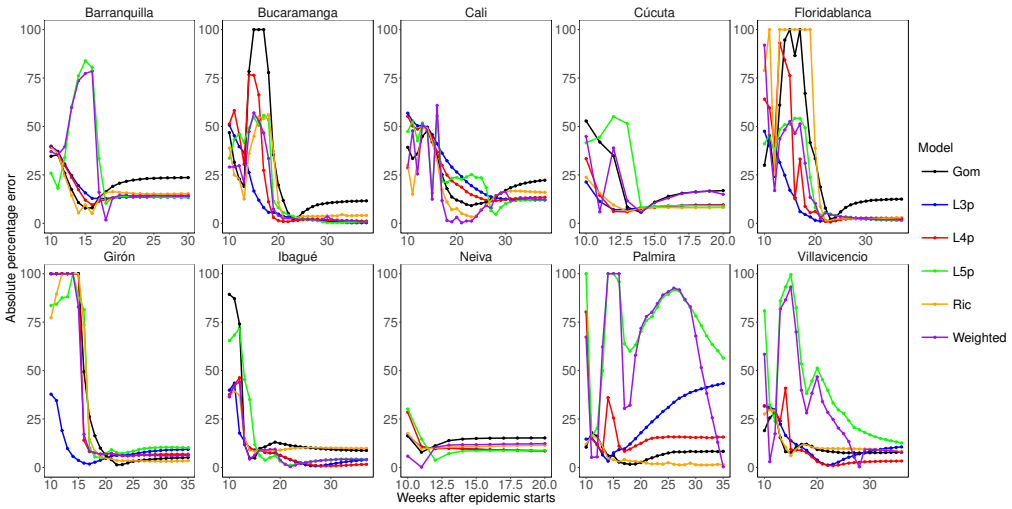


Figure 9.8: Absolute percentage error between the turning point parameter (η) and the observed turning point per city, from the Bayesian nonlinear models with Normal likelihood fitted at estimation periods 1 to t, ($t <$ T-1), and T is the oubreak length

9.5 Conclusion

In this paper we constrasted several Bayesian nonlinear model in their performance to make real-time estimation and prediction of epidemiological parameters of interest, such as the final size and the turning point of an epidemic. We chose data from the Zika outbreak in ten Colombian cities. The nonlinear models were fitted to subsequent estimation periods starting at the 10 epidemiological week after the epidemic starts per city until the last epidemiological week of the outbreak period. In addition, we produced averaged final size and turning point parameters of the epidemic per city, using weighted Bayesian parameters from the nonlinear models. The weights for the weighted parameters were obtained from the numerical optimization of the squared difference between the observed and the posterior mean predicted Zika cumulative case counts constraining the weights to add to 1.

We estimated the absolute percentage error between the observed final size and turning point and the asymptote parameter (α) and the inflection point (η) from the nonlinear models, to evaluate the estimation and prediction performance of the parameters for the final size and the turning point of the outbreak.

Real-time estimation and prediction of the final size of the epidemic were possible near to the 50% to 60% of the total outbreak length for most of the cities except Cúcuta, where the prediction was near the 75% of the outbreak length. The final size parameter from the 3PL and 4PL models, and the weighted final size were the best predictive parameters based on the generation of small absolute percentage errors during the outbreak period.

With respect to the turning point estimation, the results were similar to the final size estimation. For the turning point estimation, we relied in the estimates obtained by the Bayesian nonlinear models. The turning point parameters allowed us to estimate near the 45% to 60% of the outbreak duration, based mainly on the 3PL, 4PL and Richard models. The results contrasted with the widely used Richards model for estimation and prediction of the final size and turning point. Perhaps, the high volatility of the observed data, or the apparent symmetric distribution of the outbreak data made in general a better fit for the symmetric models (3PL and 4PL) than for the asymmetric models (5PL, Richards and 4P Gompertz).

For the study limitations, we account that the data corresponds to the addition of the confirmed cases by clinical symptoms, confirmed cases by laboratory and suspected cases, reported to the public health surveillance network of the Colombia National Institute of Health. Then, some of the cases reported and employed in this study could not be Zika cases. Zika disease clinical symptoms are similar to diseases like dengue or chikungunya, so some patients could have been reported as Zika cases in this database. In addition, underreporting could be present in the study period, but its impact is difficult to assert. The database also represent Zika cases without distinguishing the patient's gender. It is known that Zika disease is reported sligthly more in women than men [42], although similar patterns of cumulative counts are expected in women and men, allowing the use of the nonlinear models presented in the study.

The retrospective approach to the Bayesian nonlinear modeling is other limitation of

the study. Always is very easy to predict something that already happened. However, the study shows that we can start fitting the nonlinear models at early moments of the outbreak, like 10 epidemiological weeks (two and a half months) after the epidemic starts, in cities with different outbreak sizes and different starting points of the epidemic. The Bayesian nonlinear models provide an interesting approach within the epidemiological tools for estimation and prediction of epidemiological parameters.
Disease estimation and prediction using epidemiological based on phenomenological models (such as the nonlinear models employed here) could provide misleading estimation and predictions due to the unknown effects of intervention measures during the prediction periods, impacting the number of future cases, not accounted by the nonlinear model.
Zika disease is only one of many vector-borne emerging diseases ready to hit human and animal populations all around the world. We provide the Bayesian version for some well known nonlinear models based on the open source R interface to Stan (`rstan`). We found the software very robust, fast an easy to use. The Bayesian nonlinear models used mainly uniform priors for the nonlinear parameters and half-Cauchy priors for the dispersion parameters. The uniform priors were selected to avoid values producing estimation problems, like negative values in some cases or extreme negative or positive values. As future lines of study we will be testing nonlinear mixed effects models, where several cities are included in a single model, and development

9.6 References

[1] Hens N, Shkedy Z, Aerts M, Faes C, Van Damme P, Beutels P. *Modeling Infectious Disease Parameters Based on Serological and Social Contact Data A Modern Statistical Perspective*. 2012. Springer Science+Business Media, New York

[2] Van Kerkhove MD. and Ferguson NM. Epidemic and intervention modelling – a scientific rationale for policy decisions? Lessons from the 2009 influenza pandemic. *Bulletin of the World Health Organization*. 2012, **90**, 306–310.

[3] Zhou G and Yan G. Severe Acute Respiratory Syndrome Epidemic in Asia. *Emerging Infectious Diseases*. 2003, **9**, 1608–1610.

[4] Richards F. A flexible growth function for empirical use. *Journal of Experimental Botany*. 1959, **10**, 290–301.

[5] Hsieh T, Lee J, and Chang H. Sars epidemiology modeling. *Emerging Infectious Diseases*. 2004, **10**, 1165–1167.

[6] Hsieh Y. and Cheng Y. Real-time forecast of multi-wave epidemic outbreaks. *Emerging Infectious Diseases*. 2006, **12**, 122–127.

[7] Hsieh Y. Richards model: A simple procedure for real-time prediction of outbreak severity, in: Z. Ma, J. Wu, Y. Zhoue (Eds.), *Modeling and Dynamics of Infectious Diseases* (Volumen 11), 2008, Contemporary Applied Mathematics (CAM), Higher

Education Press, pp. 216–236. doi:10.1142/9789814261265_0009. http://mail.cmu.edu.tw/~hsieh/pdf/pub57.pdf

[8] Hsieh Y and Ma S. Intervention measures, turning point, and reproduction number for dengue, Singapore, 2005. *American Journal of Tropical Medicine and Hygiene*. 2009, **80**, 66–71.

[9] Hsieh Y. and Chen C. Turning points, reproduction number, and impact of climatological events for multi-wave dengue outbreaks. *Tropical Medicine and International Health*. 2009, **14**, 628–638.

[10] Hsieh Y-H, Ma S, Velasco Hernandez JX, Lee VJ, and Lim WY. Early Outbreak of 2009 Influenza A (H1N1) in Mexico Prior to Identification of pH1N1 Virus. *PLoS ONE*. 2011, **6**, e23853.

[11] Mostaço-Guidolin LC, Greer A, Sander B, Wu J, and Moghadas SM. Variability in transmissibility of the 2009 H1N1 pandemic in Canadian communities. *BMC Research Notes*. 2011, **4**, 537

[12] Wang X, Wu J, and Yang Y. Richards model revisited: Validation by and application to infection dynamics. *Journal of Theoretical Biology*. 2012, **313**, 12–19.

[13] Chan CH, Tuite AR, and Fisman DN. Historical Epidemiology of the Second Cholera Pandemic: Relevance to Present Day Disease Dynamics.. 2013, *PLoS ONE* **8**, e72498.

[14] Hsieh Y, Arazoza H, and Lounes R. Temporal trends and regional variability of 2001-2002 multiwave denv-3 epidemic in havana city: did hurricane michelle contribute to its severity? *Tropical Medicine and International Health*. 2013, **18**, 830–838.

[15] Ma J, Dushoff J, Bolker BM, and Earn DJ. Estimating Initial Epidemic Growth Rates *Bulletin Mathematical Biology*. 2014, **76**, 245–260.

[16] Hsieh Y-H. Temporal Course of 2014 Ebola Virus Disease (EVD) Outbreak in West Africa Elucidated through Morbidity and Mortality Data: A Tale of Three Countries. *PLoS ONE*. 2015, **10**, e0140810.

[17] Pinheiro J *et al.* R Core Team (2016) nlme: Linear and Nonlinear Mixed Effects Models, r package version 3.1-128. http://CRAN.R-project.org/package=nlme

[18] Gelman A, Carlin JB, Stern HS, Dunson DB, Vehtari A, Rubin DB. *Bayesian Data Analysis*. Third Edition. 2013. CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL

[19] Bolker B, Gardner B, Maunder M, et al. Strategies for fitting nonlinear ecological models in R, AD Model Builder, and BUGS *Methods in Ecology and Evolution*. 2013, **4**, 501–512.

[20] Bornkamp B. Functional Uniform Priors for Nonlinear Modeling *Biometrics* **68**. 2012, 893-901.

[21] Klauenberg K, Walzel M, Ebert B, Elster C. Informative prior distributions for ELISA analyses. *Biostatistics*. 2015, **16**, 454–464.

[22] Huson LW and Kinnersley N. Bayesian fitting of a logistic dose– response curve with numerically derived priors *Pharmaceutical Statistics*. 2009, **8**, 279-286.

[23] Lin D, Shkedy Z, Yekutieli D, Amaratunga D, Bijnens L. *Modeling Dose-Response Microarray Data in Early Drug Development Experiments Using R*, 2012, Springer-Verlag Berlin Heidelberg.

[24] Saiz J-C, Martín-Acebes MA, Bueno-Marí R, Salomón OD, Villamil-Jiménez LC, Heukelbach J, Alencar CH, Armstrong PK, Ortiga-Carvalho TM, Mendez-Otero R, Rosado-de-Castro PH and Pimentel-Coelho PM. Zika Virus: What Have We Learnt Since the Start of the Recent Epidemic? *Frontiers in Microbiology*. 2017, **8**:1554. doi: 10.3389/fmicb.2017.01554

[25] Cuevas EL, Tong VT, Rozo N, *et al.* Preliminary Report of Microcephaly Potentially Associated with Zika Virus Infection During Pregnancy - Colombia, January–November 2016. *MMWR Morb Mortal Wkly Rep*. 2016, **65**, 1409-1413.

[26] Rojas DP, Dean NE, Yang Y, et al. The epidemiology and transmissibility of Zika virus in Girardot and San Andres island, Colombia, September 2015 to January 2016. *Euro Surveillance*. 2016, **21**, 30283.

[27] Chowell G, Hincapie-Palacio D, Ospina J, et al. Using Phenomenological Models to Characterize Transmissibility and Forecast Patterns and Final Burden of Zika Epidemics. *PLOS Currents Outbreaks*. 2016, Edition 1. doi: 10.1371/currents.outbreaks.f14b2217c902f453d9320a43a35b9583.

[28] Towers S, Brauer F, Castillo-Chavez C , Falconar A, Mubayi A, Romero-Vivas C. Estimate of the reproduction number of the 2015 Zika virus outbreak in Barranquilla, Colombia, and estimation of the relative role of sexual transmission *Epidemics*. 2016, **17**, 50–55.

[29] Sebrango-Rodríguez CR, Martínez-Bello DA, Sánchez-Valdés L, Thilakarathne PJ, Del Fava E, Van Der Stuyft P, López-Quílez A, Shkedy Z. Real-time parameter estimation of Zika outbreaks using model averaging *Epidemiology & Infection*. 2017, **145**, 2313-2323.

[30] Instituto Nacional de Salud. *Reporte de notificación casos de Zika semana 32 2015 a semana 42 2016*, Instituto Nacional de Salud, Colombia 2016. [Zika notification report, week 32 2015 to week 42 2016, National Institute of Health, Colombia 2016] `http://www.ins.gov.co/Noticias/ZIKA/Forms/AllItems.aspx` [Spanish]

[31] Jorgensen M. Fitting animal growth curves. *The New Zealand Statistician*. 1981, **16**: 5–15.

[32] Pinheiro, JC, Bates, DM. *Mixed-Effects Models in S and S-PLUS*, 2000, Springer-Verlag New York, Inc.

[33] Liao J, Liu R. Re-parameterization of five-parameter logistic function. *Journal of Chemometrics*. 2009, **23**, 248–253.

[34] Polson NG and Scott JG. On the Half-Cauchy Prior for a Global Scale Parameter *Bayesian Analysis*. 2012, **7**, 887-902.

[35] Yao Y, Vehtari A, Simpson D, Gelman A. Using Stacking to Average Bayesian Predictive Distributions (with Discussion). *Bayesian Analysis*. 2018, **13**(3): 917–1003

[36] Clarke B. (2003) Comparing Bayes Model Averaging and Stacking When Model Approximation Error Cannot be Ignored *Journal of Machine Learning Research*, **4**, 683-712

[37] Stan Development Team. *RStan: the R interface to Stan*. 2016. R package version 2.14.1. `http://mc-stan.org`

[38] Carpenter B, Gelman A, Hoffman MD, Lee D, Goodrich B, Betancourt M, Brubaker MA, Guo J, Li P, Riddell A. Stan: A Probabilistic Programming Language *Journal of Statistical Software* **76**. 2017, 1–32.

[39] Hoffman MD, Gelman A. The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*. 2014, **15**, 1593–1623.

[40] Watanabe S. (2010) Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.

[41] Vehtari A, Gelman A, and Gabry J. (2016) loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models R package version 0.1.6. `https://github.com/jgabry/loo`

[42] Plourde AR, Bloch EM. A Literature Review of Zika Virus. *Emerging Infectious Diseases* 2016; **22**: 1185–1192.